International Journal of E-Health and Medical Communications
Volume 12 • Issue 5 • September-October 2021

Data Mining-Based Privacy Preservation Technique for Medical Dataset Over Horizontal Partitioned

Shivlal Mewada, Govt. Holkar (Model, Autonomous) Science College, Indore, India
https://orcid.org/0000-0001-5543-8622

ABSTRACT

The valuable information is extracted through data mining techniques. Recently, privacy preserving data mining techniques are widely adopted for securing and protecting the information and data. These techniques convert the original dataset into protected dataset through swapping, modification, and deletion functions. This technique works in two steps. In the first step, cloud computing considers a service platform to determine the optimum horizontal partitioning in given data. In this work, K-Means++ algorithm is implemented to determine the horizontal partitioning on the cloud platform without disclosing the cluster centers information. The second steps contain data protection and recover phases. In the second step, noise is incorporated in the database to maintain the privacy and semantic of the data. Moreover, the seed function is used for protecting the original databases. The effectiveness of the proposed technique is evaluated using several benchmark medical datasets. The results are evaluated using encryption time, execution time, accuracy, and f-measure parameters.

KEYWORDS

Clustering, Data Mining, Encryption, Homomorphic, K-Means++, Privacy Preservation

1. INTRODUCTION

In present time, lot of information are gathered in business houses, institutes, and government official, and the information is produced in exponential manner (Afzali & Mohammadi, 2017; Li, Lu, Choo et al, 2016). Several data mining tools are available to process the collected information and determine valuable pattern for decision making. It is seen that data mining tools also explore the various hidden information associated with individuals such as sensitive, private and confidential. Hence, several privacy preserving data mining methods are developed in literature for handling

DOI: 10.4018/IJEHMC.20210901.oa4

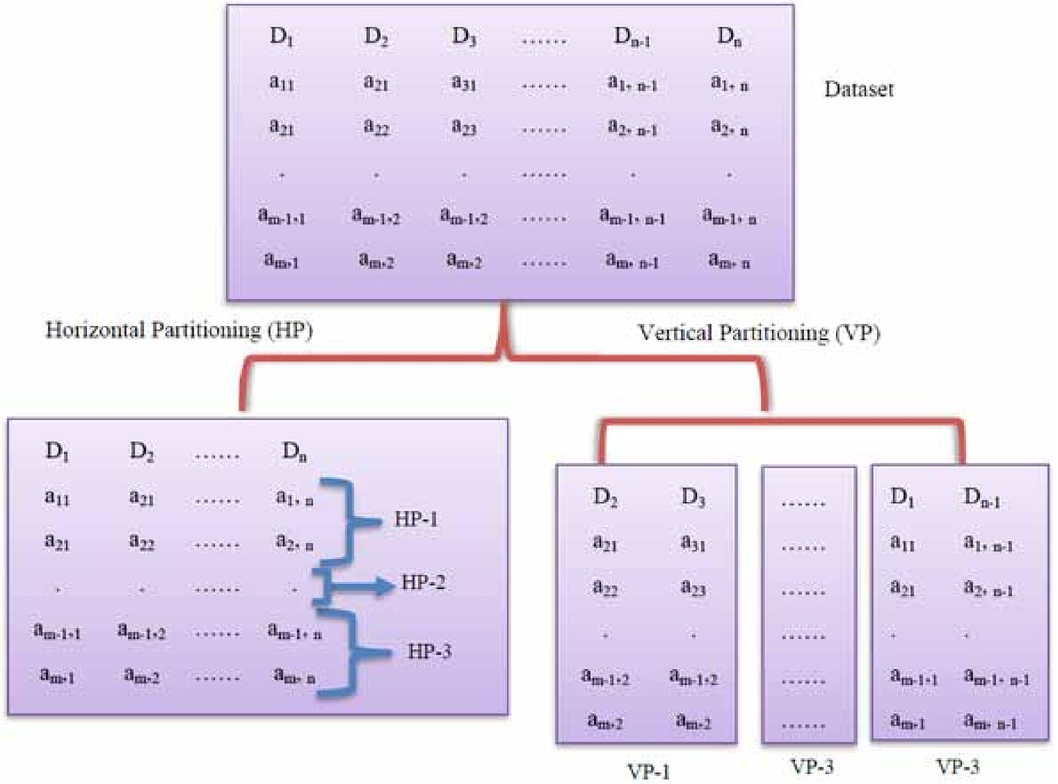
aforementioned issues (Chamikara et al., 2018; Lin et al., 2016; Mehta & Rao, 2017). These methods secure the database through perturbation technique. It is seen that PPDM methods ensure the privacy of information through converting the database. In literature, it is reported that PPDM methods provide more privacy at the mining stage and privacy can be handled during the preprocessing and postprocessing operations (Komishani et al., 2016; Upadhyay et al., 2018; Yun & Kim, 2015). It also focuses on the misuse of the sensitive of a person and organization. So, the data is modified in such a manner that intruder cannot give any comment regarding the sensitive information. In turn, the sensitive information must be protected (Dong & Pi, 2018; Qi & Zong, 2012). Other side, distributed data mining techniques become popular to extract the information in distributed resources. Several models have been designed on the concept of distributed data mining i.e. the features of collaboration of multiple parties can be used to design an effective model for avoiding the leakage of information. The distributed data mining can be either horizontal partitioned or vertical partitioned (Oliveira & Zaïane, 2007). It is interpreted as the data can be partitioned either horizontal manner or vertical manner. Figure 1 illustrates the concept of horizontal partitioning and vertical partitioning. In many applications, the sensitive information is shared among multiple parties, but privacy law stated that the information can be shared with restricted context in distributed scenarios (Mariscal et al., 2010; Matatov et al., 2010). To address the same, several distributed data mining algorithms have been developed to exchange the sensitive information among multiple parties ensuring the privacy of data (Sun et al., 2014; Zhou et al., 2015). In present time, cloud computing also attracts the attention of researchers as platform for building the distributed databases and software for third party (Ahuja et al., 2012; Grobauer et al., 2010). It is due to mobility, availability and lower cost of cloud computing. Furthermore, cloud computing can be described as clustering of multiple server design to handle the task in remote manner. Whereas, data mining can be responsible to extract the structured and consist information from unstructured and semi structured data sources. It is also noticed that distributed mining algorithms require high bandwidth of networks and having capabilities of cooperation between multiple parties (Ambulkar & Borkar, 2012). It is seen that distributed data mining algorithm having capability to handle the distributed resources in efficient manner in terms of mining and manage the data. In literature, several techniques are also reported for hiding the sensitive information especially for binary datasets (Ferrag et al., 2018; Modi et al., 2010). These techniques are the combination frequent item sets and association rules and motive of these technique is to determine the semantic information among either attributes or transactions of databases (Matatov et al., 2010). Further, to maintain the confidence and secure the sensitive information during sanitization process, these techniques can remove some of transactions and itemset from the databases. Hence, data mining and machine learning techniques have also been adopted for designing several secure protocols (Li, Yang, & Ji, 2016). Few of these are Decision tree, Bayesian networks, clustering, association rule mining and neural networks (). The objective of these techniques is to ensure the privacy of sensitive information among multiple parties during the extraction of valuable information form the datasets (Prakash & Singaravel, 2015; Sui & Li, 2017). The identification of frequent itemset and association rules are one of major issues associated with data mining-based protocol (Bhuyan & Kamila, 2015; Liu et al., 2008). It is also observed that most of data mining (DM) techniques transform the dataset in such a manner that data is not easily interpretable during the implementation of DM algorithm (González-Serrano et al., 2017; Jiang et al., 2018).

1.1 Contribution of the Work

This subsection summarizes the contributions of research work. The main contributions of this work are listed as:

1. To investigate the efficacy of horizontal partitioning of database for ensuring privacy and preserve the semantic of information over multiparty communication using cloud computing platform.

Figure 1. Illustrates the horizontal and vertical partitioning process


D1 D2 D3 …… Dn-1 Dn
a11 a21 a31 …… a1, n-1 a1, n
a21 a22 a23 …… a2, n-1 a2, n
am-1,1 am-1,2 am-1,2 …… am-1, n-1 am-1, n
am,1 am,2 am,2 …… am, n-1 am, n
Dataset
Horizontal Partitioning (HP)
Vertical Partitioning (VP)
HP-1
HP-2
HP-3
VP-1
VP-3
VP-3

2. To adopt K-Means^{++} clustering algorithm for determining the horizontal partitioning and homomorphic encryption for encrypting the information.
3. To compute the appropriate location in database for inserting the noise.
4. The performance of proposed technique is evaluated using several benchmark medical datasets.
5. Simulation results showed that proposed technique is capable to secure the sensitive information.

The organization of the paper is given as section 2 presents the related works on horizontal and vertical partitioned methods. Section 3 discusses the homomorphic encryption, K-Means^{++} clustering, data protection and recovery phases. Section 4 illustrates the results of proposed technique. Finally, the entire work is concluded in section 5.

2. RELATED WORKS

This section describes the related works on privacy preservation techniques especially, horizontal partitioned, vertical partitioned and sanitization process techniques.

Kikuchi et al. (Kikuchi et al., 2018) investigated the privacy preservation data mining techniques for epidemiological study. In this work, linear multiple regression technique is adopted for determining the relevant factors among for possible variables. The scalability of the proposed technique is evaluated using horizontal as well as vertical partitioning. To evaluate the efficacy of proposed regression technique, stroke dataset is considered. Results confirmed that horizontal partitioning method works better than vertical partitioned method.

Li et al. (Li, Lu, Choo et al, 2016) designed an efficient homomorphic encryption and a secure comparison schemes for ensuring the data privacy. Further, a cloud-based frequent itemset mining algorithm is applied for exploring the association rules. In this study, vertically partitioned dataset is considered for evaluating the efficacy of proposed schemes. Author claimed that proposed solution leaks less information than existing solutions. It is also seen that proposed solution also shares data efficiently with multiple party without compromising the data privacy.

Rong et al. (Rong et al., 2016) developed k-nearest neighbor-based privacy preservation technique for distributed database in cloud environment. Authors perform theoretical as well as experimental analysis to confirm the existence of proposed k-NN privacy preserving technique. Theoretical analysis showed that proposed technique not only preserve the privacy, but also hide access pattern in distributed database. The experimental results showed that proposed technique significantly improves the preserving rate as compared to existing technique.

Qiu et al. (Qiu, Wang, Li et al, 2017) proposed a framework based on frequent itemset mining algorithm for ensuring data privacy. The public cloud service is used to collect and encrypt the mined data. In this work, three secure frequent itemset mining protocol are developed. Authors claimed that first protocol archives higher mining performance, whereas, second protocol archives better privacy results. Hence, it is stated that proposed framework Qiu et al. (Qiu, Xu, Ahmad et al, 2017) developed novel protocol for ensuring privacy-preserving. This protocol considers linear regression and applies on horizontally partitioned data. Authors claimed that proposed protocol architecture supports multiple clients and also includes two non-colluding servers. It is seen that clients can submit the encrypted data to the server and further, two servers determine the regression model collaboratively. Paillier homomorphic encryption and data masking technique are associated with protocol. The simulation results ensured that proposed protocol is an efficient approach with negligible errors.

Pang and Wang (Pang & Wang, 2020) developed novel homomorphic cryptosystem that supports multiple cloud users with different public keys. In this work, secure association rule mining scheme is used with outsourced databases from multiple parties in a twin-cloud architecture. Simulation results showed that proposed cryptosystem is reasonable significant than other models.

Skarkala et al. (Skarkala et al., n.d.) introduced a privacy preserving data mining algorithm for horizontal and vertical partitioned databases. This work considers the multi candidate election scheme for building the naïve bayes classifier. The experimental results showed that proposed algorithm ensures the privacy of data through mining process.

Sekhavat (Sekhavat, 2020) analyzed the frequent itemset mining protocol for privacy purpose and developed a new protocol collusion-free model, called CFM. CFM is applied on horizontal partitioned database. Apart from these, a new secret sharing and summation is employed in CFM. The efficacy of proposed CFM is evaluated using disclosure of sensitive information.

Lekshmy and Rahiman (Lekshmy & Rahiman, 2020) developed a kernel K-Means based privacy preserving data mining model for distributed databases. Authors consider horizontal as well as vertical partitioned method for ensuring the data privacy. In sanitization process, artificial bee colony-based algorithm is applied for generating the best key value for secure data. The performance of proposed model is evaluated using clustering accuracy, processing time and data transmission time. It is stated that proposed model achieves better hacking results as compared to existing models.

Li et al. (Li et al., 2019) preserved the privacy of the data using an outsourced privacy-preserving C4.5 algorithm. The proposed technique works with horizontal and vertical partitioned data for multiple parties. Further, two protocols are designed for ensuring the data. These are outsourced privacy preserving weighted average protocol (OPPWAP) and outsourced secure set intersection protocol. Authors claimed that proposed model performs less computation on client side.

Liu et al. (Liu et al., 2018) introduced a hybrid privacy-preserving clinical decision support system in fog-cloud computing, called HPCS. The proposed system consists of a lightweight data mining method for secure monitoring the health condition of patients. An outsourced inner-product protocol is applied on fog layer for securing single layer neural network. Whereas, on cloud layer, piecewise polynomial calculation privacy preserving protocol is applied for securing activation function of multiple layer neural network. Simulation results showed that proposed system is capable to overcome the privacy leakage problem with authorized parties.

Omer et al. (Omer et al., 2017) designed a distributed privacy preserving protocol with multiple missing imputation data. The missing data is handled through multivariate imputation by chained equations and further, privacy is maintained through Paillier cryptosystem. Finally, SVM technique is introduced as semi-honest approach over vertical partition of the data. Simulation results showed that proposed protocol achieves better processing time as compared to existing model especially for distributed database.

Rajesh and Selvakumar (Rajesh & Selvakumar, 2019) developed a privacy preservation data-mining scheme with data-mining perturbation merged approach for ensuring the data privacy. In this work, association rules with cryptography technique are used for data privacy and preservation. Further, capabilities of neural network and deep learning model is also investigated for prediction medical datasets. Authors claimed that proposed privacy algorithm maintains the confidentiality of information.

Tran et al. (Tran et al., 2020) developed a new secure data analysis technique, called SmartClass using garbled circuit technique. The encryption step consists of additive homomorphism and binary Elliptic Curve Cryptography (ECC) algorithm for encryption of data. Authors claimed that proposed technique achieves promising results than others.

Wu et al. (Wu et al., 2019) proposed a density-based clustering algorithm for securing the sensitive information. The proposed density-based clustering algorithm implements the multi objective particle swarm optimization algorithm, called CMPSO. The proposed CMPSO algorithm works according the user preference and performance of proposed algorithm is evaluated using two benchmark datasets. The simulation results stated that proposed CMPSO provides better results than single objective algorithms.

Shailaja and Rao (Shailaja & Rao, 2019) presented a novel privacy preserving data mining technique for securing the privacy of data. The proposed technique works in two phase-sanitization and restoration. Prior to proceed with phases, some association rules are extracted for the datasets. Further, opposition intensity-based cuckoo search algorithm is adopted for generating the optimal key for hiding the information. In this work, four research issues are considered to evaluate the performance of proposed technique such as hiding failure rate, information preservation rate, and false rule generation, and degree of modification. It is observed that proposed technique minimizes all four issues.

Sheela and Vijayalakshmi (Sheela & Vijayalakshmi, 2017) developed the partition-based perturbation technique for ensuring the privacy of data in distributed databases. The proposed technique is applied on vertical partitioned dataset. Furthermore, the data is recursively partitioned between various parties using third party coordinator. Its showed that proposed technique ensures the statistical technique relationship between attributes.

Domadiya and Rao (Domadiya & Rao, 2019) presented an association rule mining-based privacy preservation technique for ensuring the privacy in healthcare dataset. The medical examination and outpatient data of Taiwan is considered for evaluating the performance of proposed approach. It is observed that association rule mining scheme is applied to partition the data vertically. The efficacy of proposed approach is evaluated using privacy preservation, communication and computation cost. Authors claimed that proposed approach is better than others in terms of theoretical and experimental analysis.

Yu et al. (Yunhong et al., 2009) designed an efficient and secure privacy preserving technique based on support vector machine (SVM). This study considers the vertically partitioned data. It is seen that proposed technique shares the information between multiple parties without revealing the data and classification information.

Xia et al. (Xia et al., 2019) introduced a novel differential privacy mechanism to preserve the privacy in heterogeneous databases. Authors consider the vertical partitioned data for securing and preserving the information among multiple parties. The information-based method is used to determine the dependencies between attributes and privacy labels. Moreover, Laplace method is adopted for maintain the privacy in heterogenous environment. Several benchmark datasets are considered for examining the efficiency of proposed mechanism. It is noticed that proposed method achieves balanced privacy and utility.

Oliveiraa and Zaıaneb (Oliveira & Zaïane, 2007) developed a new method for privacy-preserving clustering called Dimensionality Reduction-Based Transformation (DRBT). Authors consider both centralized and vertical partitioned method to implement the DRBT method. Furthermore, three issues are considered security, communication cost, and accuracy to evaluate the performance of proposed method. It is noticed that DRBT method successfully preserve the semantic information of the data.

2.1 Technical Gaps

In this study, twenty-one research papers are considered to determine the existing research gaps in the field of data privacy preservation especially for medical dataset. Though existing studies, it is found that both horizontal and vertical methods are used for partitioned the datasets for preserving the privacy of data. It is also observed that horizontal method having advantage over vertical partitioning methods. So, in this work, we considered the horizontal partitioning method for partitioning of datasets for privacy preservation. The other point regarding the horizontal partitioned method is less computationally extensive than vertical and other privacy preservation methods. It is also observed that K-Means++ is one of effective and efficient method for clustering. Hence, in this work, K-Means++ algorithm is adopted for determining the effective horizontal partitioning. In spite this, homomorphic scheme is applied for encryption of data. In this scheme, RSA algorithm is considered for effective encryption of data and finally, a data protection phase is also used to recover the encrypted data.

3. PROPOSED K-MEANS^{++} BASED PRIVACY ALGORITHM

In this section, K-Means^{++} clustering-based privacy preserving algorithm is discussed. The proposed algorithm works in two phases- i) Data Protection Phase, and ii) Data Recovery Phase. The data protection phase corresponds for adding the noise in original dataset and converts into protected dataset. Whereas, the added noise is detected and removed in data recovery phase and obtained original dataset from protected dataset. First, we discuss the homographic encryption in subsection 4.1 and further, detailed description of the proposed K-Means^{++} clustering-based privacy preserving algorithm is presented in subsection 4.2.

3.1 Homomorphic Encryption

In homomorphic encryption, several computations are performed on encrypted data (Lekshmy & Rahiman, 2020; Qiu, Wang, Li et al, 2017; Qiu, Xu, Ahmad et al, 2017). These computations are similar as ciphertext operations i.e. to convert the plain test into cipher text. allows certain calculations of encrypted data; this is equivalent to ciphertext operations after the encryption of a plaintext operation. The well-known example of homomorphic encryption is RSA algorithm (Rivest & Silverman, 1999). This algorithm performs fast encryption and decryption as compared to same class of algorithms. The RSA algorithm is more reliable and secure, because it considers the large integer decomposition, in turn, longer the RSA key, more secure the algorithm. The computation steps of RSA algorithm are mentioned in Algorithm 1.

Algorithm 1. Steps of RSA Algorithm

Step 1:	Select two large prime numbers g and h than can satisfy the following equation: $\gcd\left(gh,\left(g-1\right)\left(h-1\right)\right)=1$ (1)
Step 2:	Compute $\delta = g \times h \ \ and \ \ \lambda = lcm\left(g-1,h-1\right)$.
Step 3:	Choose an integer $b \in Q_{n^2}$ in random order and check the following: $\exists t = 1 / \left(L\left(b^{\lambda}, \bmod \delta^{2}\right)\right) \times \bmod \delta, L\left(\mu\right) = \left(\mu - 1\right) / n$ (2) Then, plaintext(pt) is (δ ,b) and ciphertext(ct) is $\left(\lambda, \mu\right)$.
Step 4:	For the encryption process $E_{pt}\left(v,w\right) \rightarrow y;$ choose $w \in Q_{\delta}^{*}$ for text v and ciphertext is $y = b^{w} \times r^{\delta} \times \bmod \delta^{2}$.
Step 5:	For the decryption process $D_{ct}\left(y\right) \rightarrow v;$ the message (v)is decrypted using $L\left(y^{\lambda} \times \bmod \delta^{2}\right) \times \mu \times \bmod \delta$.

3.2 K-Means^{++} Clustering Algorithm

This subsection describes the K-means^{++} clustering algorithm. The K-Means^{++} algorithm is an improved variant of K-Means algorithm (Arthur & Vassilvitskii, 2006a). In literature, it is reported that several shortcomings are associated with K-Means algorithm such as undesired and weak clustering results, and initial selection of clusters (Arthur & Vassilvitskii, 2006b; Arthur & Vassilvitskii, 2006c). This algorithm chooses an initial cluster center (suppose, c_1) from the given dataset in uniform order. The rest of cluster centers are computed using a probability function instead of random initialization and further, proceeded with K-Means algorithm steps. The computational procedure of K-Means^{++} clustering algorithm is presented in Algorithm 1. Figure 2 demonstrates the working of the proposed privacy preservation K-Means^{++} technique for protecting the privacy leakage.

3.2.1 Data Protection Phase

This phase corresponds to divide the original data into k clusters using K-Means^{++} algorithm. In turn, the original dataset is partitioned into K-partitions. These partitions are achieved through-Means^{++} clustering algorithm. It is stated that protected dataset is similar to original dataset and information stored into original dataset can be preserved in protected dataset. Next step is to choose any partition (select one cluster (c) out of K clusters) in random number. Further, a farthest data instance (d) is computed through the selected centroid (c) and determine the set of noise using distance between (c, d), and offset ratio $\left(\alpha\right)$. Finally, the noise is added into original dataset to generate protected dataset. The location for adding the noise in dataset is computed using equation 6. This work also considers traditional and progressive noise generation schemes for generating the noise. $\left(\left|k\right| \times r\right)$ noises are generated through traditional noise scheme based on distance between (c, d), and offset ratio $\left(\alpha\right)$. Whereas, progressive scheme computes the first noise using c, d and α with respect to selected data d. The second noise is generated through the noise associated with data d. The above process is repeated, until $\left(\left|k\right| \times r\right)$ noises are not generated. It is also stated that for preserving the information

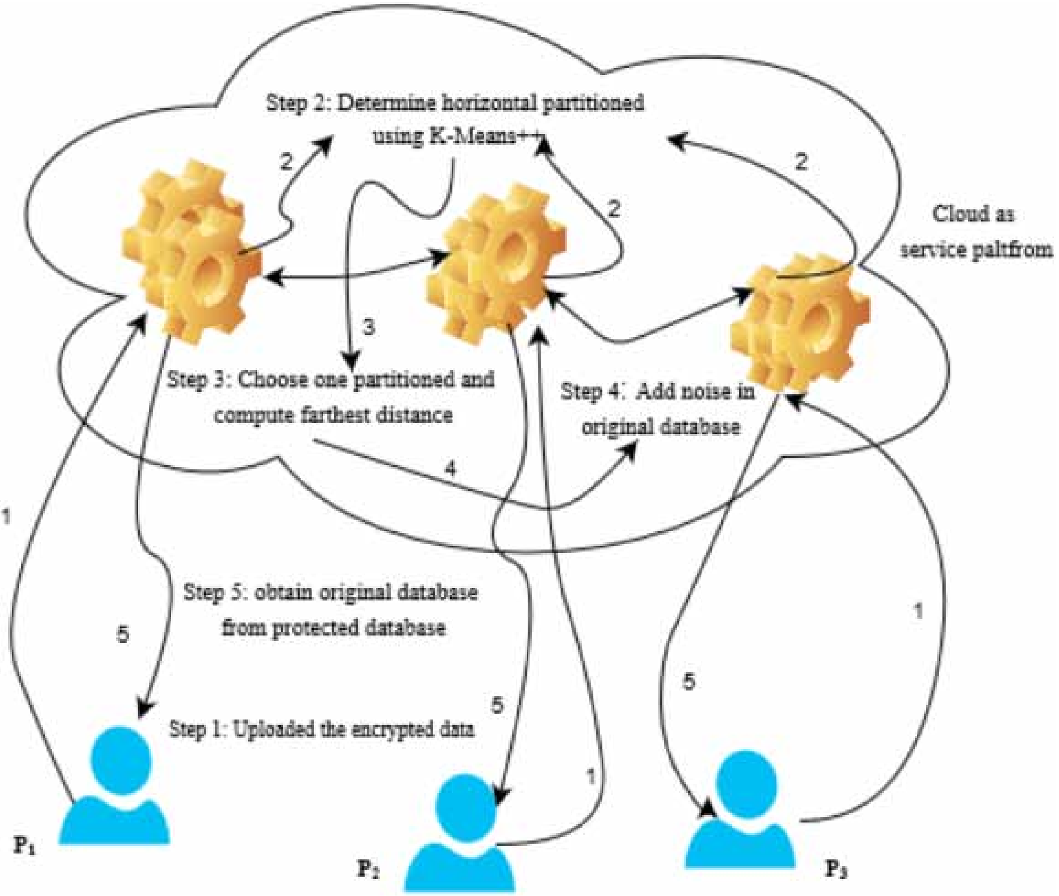
Algorithm 2. Steps of K-Means^{++} clustering Algorithm

Input:	A dataset (D) and number of clusters (K)
Output:	Partitioned Dataset (D') into Cluster(K)
Step 1:	Load the dataset (D) and determine initial cluster center ($c_1 \in K$) from dataset(D) in uniform order.
Step 2:	Compute the next cluster center (c_i) such that $\left(c_i\right) = x' \in D$ using a probability function mentioned in equation 3: $c_i = \frac{dist\left(x'\right)^2}{\sum_{x \in D} dist\left(x\right)^2}$ (3) dist(x)denotes the shortest distance between data (x) to nearest center that randomly chosen.
Step 3:	Repeat the step 2, until all cluster centers are not determined.
Step 4:	Compute the Euclidean distance between cluster centers ($c_i \in K$) and each data (x) presented in dataset (D).
Step 5:	Allocate the data(x) to clusters ($c_i \in K$) with minimum Euclidean distance.
Step 6:	Recompute the new clusters using equation 4: $c_{i,new} = \left(1 / c_i\right) \sum_{k=1}^{c_i} x_i$ (4)
Step 7:	Repeat the steps 4-6, until, there is no change in data allocation between clusters.

Algorithm 3. K-Means^{++} clustering-based data protection process

Input:	A dataset (D), number of clusters (K), noise ratio (r), offset ratio $\left(\alpha\right)$, and seed (s)
Output:	Protected Dataset (D')
Step 1:	Partitioned the dataset (D) into K-cluster using K-Means^{++} clustering algorithm.
Step 2:	Choose one cluster k from K-clusters such that $k \in K$.
Step 3:	Compute the farthest data (d) within cluster from cluster centroid (c) and determine the set of noises $\left(N\right) = \left\{n_1, n_2, n_3, \ldots, n_{\lvert N \rvert}\right\}\left(\left\lvert N\right\rvert = \left\lvert D\right\rvert \times r\right)$ using equation 5: $n_i^u = d^u + \alpha \times \left(distance\left(c, d\right)\right)$ (5)
Step 4:	Add noise $\left(n_i\right)$ on position $\left(p_i\right)$ in the given dataset (D) using equation 6 to obtain protected dataset. $p_i = \left\lvert D\right\rvert \times rand\left(s\right)$ (6) Protected dataset can be described as $D' = D \cup N\left(\left\lvert D'\right\rvert = \left\lvert D\right\rvert + \left\lvert N\right\rvert\right)$

Figure 2. Demonstrates the working of the proposed privacy preservation K-Means^{++} technique


Step 2: Determine horizontal partitioned using K-Means++
2
2
2
Cloud as service paltfrom
3
Step 3: Choose one partitioned and compute farthest distance
Step 4: Add noise in original database
4
1
Step 5: obtain original database from protected database
5
1
5
5
Step 1: Uploaded the encrypted data
1
P₁
P₂
P₃

(maximum knowledge) in original dataset and also achieving protected dataset similar to original dataset, the proposed K-Means++ based privacy preservation algorithm can only add the noise for privacy protection. It can be accomplished through seed (s). It determines the location for inserting the noise in original dataset. The algorithmic step of the data protection phase is listed in Algorithm 3.

3.2.2 Data Recovery Phase

This subsection describes the data recovery process i.e. to determine the original dataset (D) from protected dataset (D'). So, to recover the original dataset, first noise locations in protected dataset are identified using equation 2 and determine the noises. Furthermore, the detected noises are deleted from the dataset and obtains the original dataset. Here, it is also specified that the noise locations are determined through seed (s) from the protected dataset (D') it is also mentioned that seed (s) is similar to a cryptographic key and it must be protected. The steps of data recovery phase are mentioned in Algorithm 4.

Algorithm 4. Steps of data recovery process

Input:	Protected Dataset (D’), number of data in dataset (D), number of clusters (K), and seed (s)
Output:	Original Dataset (D)
Step 1:	Compute the position $\left(p_i\right)$ of noise $\left(n_i\right)$ in protected dataset (D’) using equation 2 and delete the noise $\left(n_i\right)$.
Step 2:	Retrieve the original dataset (D).

4. EXPERIMENTAL RESULTS

This section discusses the effectiveness of the proposed privacy preservation algorithm. The proposed privacy algorithm is implemented in Java language using CPU Intel Core i5, RAM 8 GB, and Microsoft Windows 10 operating system. Several medical datasets are considered for evaluating the performance of proposed algorithm. These algorithms are downloaded from the UCI repository and details of these datasets are presented in Table 1. These datasets are autism- adult, autism-child, cryotherapy, immunotherapy, heart and WPBC. The simulation results are evaluated using encryption time, execution time, accuracy and f-measure parameters. The performance of proposed privacy preservation technique is compared with, DPC, K-means, PPDC and SC techniques.

4.1 Relation of Data Size and Encryption Time

This section explores the relation between data size and encryption time. It is sated that distance between each sample is computed prior to clusters the dataset using cloud. To determine the relation between data size and encryption time, we divide the dataset into four parts i.e. first part corresponds to one fourth data instance, second part corresponds to one half data instance, third part corresponds to third fourth data instance and fourth part consider the entire dataset. Figure 3 shows the relation between the different data size with respect to encryption time. It is analyzed that encryption time for immune-therapy dataset varies in the range of 18 seconds to 54 seconds, for the Cryo-therapy dataset, encryption time is ranging in between 12 seconds to 48 seconds. Whereas, heart dataset requires 54 to 134 second for encrypting the entire dataset. The encryption time range for Autism-Child dataset is ranging in between 48 to 131 seconds. The encryption time for Autism-adult dataset is 64 second to 208 seconds. While, WPBC dataset taking 59 seconds to 184 seconds for encryption the information. Hence, it is observed that size of data has significant impact on the encryption time. The encryption time increases as the size of dataset is increased. The execution time of the proposed

Table 1. Detailed description of datasets used in this study (UCI, n.d.)

Sr. No.	Dataset	No. of Samples	No. of Attributes	No. of Clusters
1	WPBC	198	33	2
2	Heart	303	13	2
3	Autism-Adult	704	21	2
4	Autism-Child	292	21	2
5	Cryotherapy	90	7	2
6	Immunotherapy	90	8	2

Figure 3. Illustrates the impact of different data size on encryption time

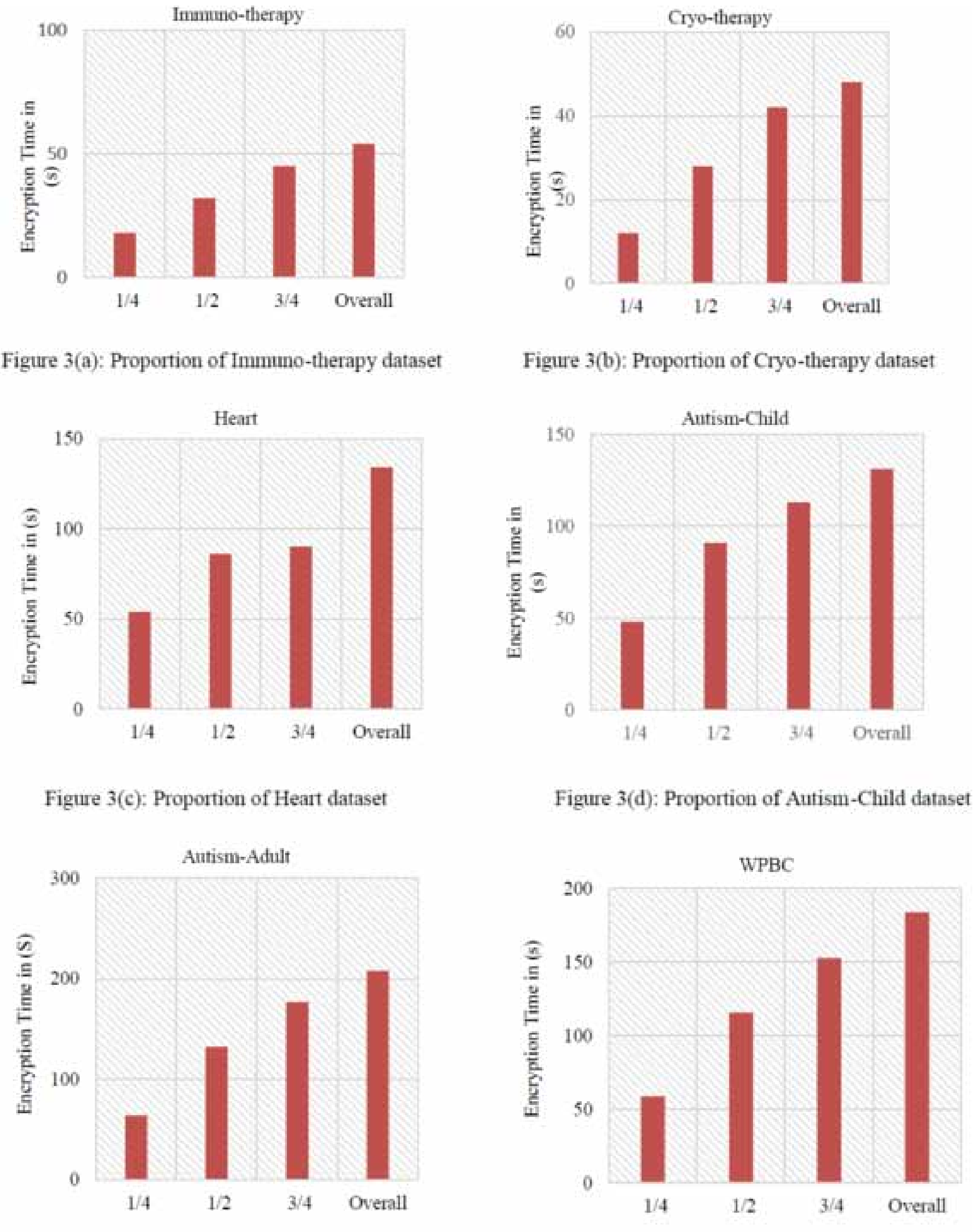


Figure 3(a): Proportion of Immuno-therapy dataset

Figure 3(b): Proportion of Cryo-therapy dataset

Figure 3(c): Proportion of Heart dataset

Figure 3(d): Proportion of Autism-Child dataset

Figure 3(e): Proportion of Autism-Adult dataset

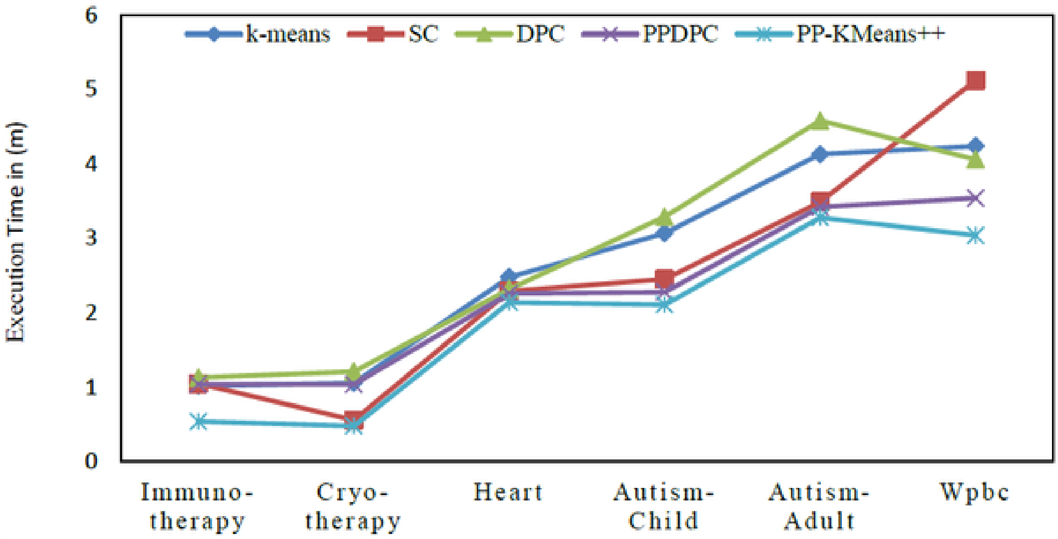
Figure 3(f): Proportion of WPBC dataset

privacy preserving K-Means^{++} (PP-K-Means^{++}) is also compared with state of art existing studies. These studies include different algorithms such as K-Means, SC, DPC and PPDPC. On the analysis, it's concluded that proposed PP-K-Means^{++} technique takes less time for executing the algorithm.

Table 2. Performance comparison of different algorithm using execution time (m) parameter

Datasets	k-means	SC	DPC	PPDPC	PP-K-Means^{++}
Immuno-therapy	1.02	1.05	1.13	1.04	0.54
Cryo-therapy	1.06	0.56	1.21	1.04	0.48
Heart	2.48	2.29	2.32	2.26	2.14
Autism-Child	3.07	2.45	3.29	2.28	2.11
Autism-Adult	4.13	3.49	4.58	3.42	3.28
WPBC	4.24	5.12	4.06	3.54	3.04

Figure 4. Execution tine of different algorithms using all dataset used in the study



In this work, data is horizontal partitioned. The horizonal partitioning is achieved by determining the optimum number of clusters in data. It is also stated that proposed K-Means^{++} algorithm is less sensitive to initial cluster center, in turn, less chance to stuck in local optima and converge on global optimal solution. Hence, the proposed PP-K-Means^{++} technique converge on minimum running time as compared to other techniques.

Figure 4 demonstrates the comparison of execution time of different algorithm using all datasets considered in this study. It is seen that proposed PP-K-Means^{++} algorithm archives minimum time for all datasets. Whereas, DPC algorithm converges on maximum time with most of datasets as compared to rest of algorithms. It is also seen that SC technique converges on maximum time for WPBC dataset. Hence, it is stated that proposed PP-K-Means++ is an effective technique for preserving and protecting the information.

4.2 Clustering Performance Evaluation

This subsection illustrates the performance of algorithms in terms of clustering of dataset (horizontal partitioned). The performance is evaluated using accuracy and f-measure parameters. The clustering results of proposed K-Means++ and all other algorithms are reported in Table 3. It is seen that

Table 3. Performance comparison of different clustering using accuracy and f-measure parameters

Dataset	Algorithms	Accuracy	F-Measure	Dataset	Algorithms	Accuracy	F-Measure
Immuno-therapy	k-means	0.783	0.767	WPBC	k-means	0.657	0.618
	SC	0.749	0.724		SC	0.643	0.637
	DPC	0.776	0.732		DPC	0.692	0.668
	PPDPC	0.781	0.753		PPDPC	0.721	0.696
	K-Means^{++}	**0.809**	**0.781**		K-Means^{++}	**0.768**	**0.743**
Heart	k-means	0.706	0.69	Autism-Child	k-means	0.834	0.826
	SC	0.728	0.698		SC	0.821	0.817
	DPC	0.734	0.726		DPC	0.825	0.804
	PPDPC	0.758	0.718		PPDPC	0.847	0.829
	K-Means^{++}	**0.774**	**0.762**		K-Means^{++}	**0.861**	**0.848**
Cryo-therapy	k-means	0.817	0.795	Autism-Adult	k-means	0.726	0.713
	SC	0.802	0.778		SC	0.687	0.674
	DPC	0.826	**0.817**		DPC	0.735	0.721
	PPDPC	0.813	0.782		PPDPC	**0.756**	0.742
	K-Means^{++}	**0.832**	0.812		K-Means^{++}	0.754	**0.749**

K-Means^{++} algorithm achieves better accuracy rate for all datasets except autism-adult dataset. The accuracy rate of K-Means^{++} algorithm is 0.809, 0.768, 0.774, 0.816 and 0.832 for immunotherapy, WPBC, heart, autism-child and cryotherapy datasets respectively. For autism adult dataset, PPDC algorithm achieves better accuracy rate i.e. 0.756 than other algorithms, but the K-Means^{++} also obtains competitive accuracy rate i.e. 0.754. On the analysis of f-measure parameter, it is reported that K-Means^{++} archives higher f-measure rate with most of datasets except cryotherapy dataset. K-Means^{++} obtains 0.781, 0.743, 0.762, 0.848 and 0.749 for immunotherapy, WPBC, heart, autism-child and autism-adult datasets respectively. It is observed that DPC technique archives better f-measure rate i.e. 0.817 for cryotherapy dataset as compared to rest of algorithms. The f-measure rate (0.812) of K-Means^{++} algorithm is nearly equal to DPC algorithm. Hence, it is concluded that proposed algorithm obtains better horizontal partitioned with most of datasets.

5. CONCLUSION

This paper presents a distributed privacy preserving technique based on K-Means^{++} algorithm. The aim of the proposed technique is to address the privacy leakage. Furthermore, this work considers the cloud computing as a service platform for collecting the data and sharing between multiple parties. Moreover, it is said that K-Means^{++} clustering algorithm is implemented on cloud computing platform and aim of this algorithm is to determine the optimal horizontal partitioning for ensuring the privacy of information. The data protection phase is applied to add noise in the original dataset and converted into protected dataset. The original database is retrieved from protected dataset using determining the location of noise. The performance of proposed privacy algorithm is evaluated using six benchmark medical datasets. The encryption and execution times are considered to examine the efficacy of proposed algorithm. The results are compared with state of art privacy preserving technique in literature. It is observed that proposed privacy preserving technique more optimum results than others. The clustering results of the proposed technique are also compared with other techniques using accuracy and f-measure parameters. It is showed that proposed clustering algorithm more accurate results than other techniques. In future, some meta heuristic techniques will be adopted for effective horizontal partitioning of the datasets. The efficacy of vertical partition for effective data privacy can be explored. Furthermore, privacy preservation scheme can be implemented in distributed environment.

REFERENCES

Afzali, G. A., & Mohammadi, S. (2017). Privacy preserving big data mining: Association rule hiding using fuzzy logic approach. *IET Information Security*, *12*(1), 15–24. doi:10.1049/iet-ifs.2015.0545

Ahuja, S. P., Mani, S., & Zambrano, J. (2012). A survey of the state of cloud computing in healthcare. *Network and Communication Technologies*, *1*(2), 12. doi:10.5539/nct.v1n2p12

Ambulkar, B., & Borkar, V. (2012, April). Data mining in cloud computing. In *MPGI National Multi Conference* (*Vol. 2012*). Academic Press.

Anjum, A., Ahmed, T., Khan, A., Ahmad, N., Ahmad, M., Asif, M., Reddy, A. G., Saba, T., & Farooq, N. (2018). Privacy preserving data by conceptualizing smart cities using MIDR-Angelization. *Sustainable Cities and Society*, *40*, 326–334. doi:10.1016/j.scs.2018.04.014

Arthur, D., & Vassilvitskii, S. (2006a). k-means++: The advantages of careful seeding. Stanford.

Arthur, D., & Vassilvitskii, S. (2006b, June). How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry* (pp. 144-153). doi:10.1145/1137856.1137880

Arthur, D., & Vassilvitskii, S. (2006c, October). Worst-case and smoothed analysis of the ICP algorithm, with an application to the k-means method. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)* (pp. 153-164). IEEE. doi:10.1109/FOCS.2006.79

Bennati, S., & Pournaras, E. (2018). Privacy-enhancing aggregation of Internet of Things data via sensors grouping. *Sustainable Cities and Society*, *39*, 387–400. doi:10.1016/j.scs.2018.02.013

Bhuyan, H. K., & Kamila, N. K. (2015). Privacy preserving sub-feature selection in distributed data mining. *Applied Soft Computing*, *36*, 552–569. doi:10.1016/j.asoc.2015.06.060

Chamikara, M. A. P., Bertók, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. *Pervasive and Mobile Computing*, *48*, 1–19. doi:10.1016/j.pmcj.2018.05.003

Domadiya, N., & Rao, U. P. (2019). Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. *Procedia Computer Science*, *148*, 303–312. doi:10.1016/j.procs.2019.01.023

Dong, Y., & Pi, D. (2018). Novel privacy-preserving algorithm based on frequent path for trajectory data publishing. *Knowledge-Based Systems*, *148*, 55–65. doi:10.1016/j.knosys.2018.01.007

Ferrag, M. A., Maglaras, L. A., Janicke, H., Jiang, J., & Shu, L. (2018). A systematic review of data protection and privacy preservation schemes for smart grid communications. *Sustainable Cities and Society*, *38*, 806–835. doi:10.1016/j.scs.2017.12.041

González-Serrano, F. J., Navia-Vázquez, Á., & Amor-Martín, A. (2017). Training support vector machines with privacy-protected data. *Pattern Recognition*, *72*, 93–107. doi:10.1016/j.patcog.2017.06.016

Grobauer, B., Walloschek, T., & Stocker, E. (2010). Understanding cloud computing vulnerabilities. *IEEE Security and Privacy*, *9*(2), 50–57. doi:10.1109/MSP.2010.115

Jiang, R., Lu, R., & Choo, K. K. R. (2018). Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. *Future Generation Computer Systems*, *78*, 392–401. doi:10.1016/j.future.2016.05.005

Kikuchi, H., Hamanaga, C., Yasunaga, H., Matsui, H., Hashimoto, H., & Fan, C. I. (2018). Privacy-preserving multiple linear regression of vertically partitioned real medical datasets. *Journal of Information Processing*, *26*(0), 638–647. doi:10.2197/ipsjjip.26.638

Komishani, E. G., Abadi, M., & Deldar, F. (2016). PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. *Knowledge-Based Systems*, *94*, 43–59. doi:10.1016/j.knosys.2015.11.007

Lekshmy, P. L., & Rahiman, M. A. (2020). A sanitization approach for privacy preserving data mining on social distributed environment. *Journal of Ambient Intelligence and Humanized Computing*, *11*(7), 2761–2777. doi:10.1007/s12652-019-01335-w

Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016). Privacy-preserving-outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1847–1861. doi:10.1109/TIFS.2016.2561241

Li, Y., Jiang, Z. L., Yao, L., Wang, X., Yiu, S. M., & Huang, Z. (2019). Outsourced privacy-preserving C4. 5 decision tree algorithms over horizontally and vertically partitioned dataset among multiple parties. *Cluster Computing*, *22*(1), 1581–1593. doi:10.1007/s10586-017-1019-9

Li, Y., Yang, J., & Ji, W. (2016). Local learning-based feature weighting with privacy preservation. *Neurocomputing*, *174*, 1107–1115. doi:10.1016/j.neucom.2015.10.038

Lin, J. C. W., Wu, T. Y., Fournier-Viger, P., Lin, G., Zhan, J., & Voznak, M. (2016). Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining. *Engineering Applications of Artificial Intelligence*, *55*, 269–284. doi:10.1016/j.engappai.2016.07.003

Liu, L., Kantarcioglu, M., & Thuraisingham, B. (2008). The applicability of the perturbation-based privacy preserving data mining for real-world data. *Data & Knowledge Engineering*, *65*(1), 5–21. doi:10.1016/j.datak.2007.06.011

Liu, X., Deng, R. H., Yang, Y., Tran, H. N., & Zhong, S. (2018). Hybrid privacy-preserving clinical decision support system in fog–cloud computing. *Future Generation Computer Systems*, *78*, 825–837. doi:10.1016/j.future.2017.03.018

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. doi:10.1017/S0269888910000032

Matatov, N., Rokach, L., & Maimon, O. (2010). Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences*, *180*(14), 2696–2720. doi:10.1016/j.ins.2010.03.011

Mehta, B. B., & Rao, U. P. (2017). Privacy preserving big data publishing: A scalable k-anonymization approach using MapReduce. *IET Software*, *11*(5), 271–276. doi:10.1049/iet-sen.2016.0264

Modi, C., Rao, U. P., & Patel, D. R. (2010, March). A survey on preserving privacy for sensitive association rules in databases. In *International Conference on Business Administration and Information Processing* (pp. 538-544). Springer. doi:10.1007/978-3-642-12214-9_96

Oliveira, S. R., & Zaïane, O. R. (2007). A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. *Computers & Security*, *26*(1), 81–93. doi:10.1016/j.cose.2006.08.003

Omer, M. Z., Gao, H., & Mustafa, N. (2017). Privacy-preserving of SVM over vertically partitioned with imputing missing data. *Distributed and Parallel Databases*, *35*(3-4), 363–382. doi:10.1007/s10619-017-7203-3

Pang, H., & Wang, B. (2020). Privacy-Preserving Association Rule Mining Using Homomorphic Encryption in a Multikey Environment. *IEEE Systems Journal*, 1–11. doi:10.1109/JSYST.2020.3001316

Prakash, M., & Singaravel, G. (2015). An approach for prevention of privacy breach and information leakage in sensitive data mining. *Computers & Electrical Engineering*, *45*, 134–140. doi:10.1016/j.compeleceng.2015.01.016

Qi, X., & Zong, M. (2012). An overview of privacy preserving data mining. *Procedia Environmental Sciences*, *12*, 1341–1347. doi:10.1016/j.proenv.2012.01.432

Qiu, S., Wang, B., Li, M., Liu, J., & Shi, Y. (2017). Toward practical privacy-preserving frequent itemset mining on encrypted cloud data. IEEE Transactions on Cloud Computing.

Qiu, S., Xu, G., Ahmad, H., & Wang, L. (2017). A robust mutual authentication scheme based on elliptic curve cryptography for telecare medical information systems. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 7452–7463. doi:10.1109/ACCESS.2017.2780124

Rajesh, N., & Selvakumar, A. A. L. (2019). Association rules and deep learning for cryptographic algorithm in privacy preserving data mining. *Cluster Computing*, *22*(1), 119–131. doi:10.1007/s10586-018-1827-6

Rivest, R. L., & Silverman, R. D. (1999, November). Are Strong Primes Needed for RSA? *1997 RSA laboratories seminar series, seminars proceedings*.

Rong, H., Wang, H. M., Liu, J., & Xian, M. (2016). Privacy-preserving k-nearest neighbor computation in multiple cloud environments. *IEEE Access: Practical Innovations, Open Solutions*, *4*, 9589–9603. doi:10.1109/ACCESS.2016.2633544

Sekhavat, Y. A. (2020). CFM: Collusion-free model of privacy preserving frequent itemset mining. *International Journal of Information and Computer Security*, *13*(3-4), 249–267. doi:10.1504/IJICS.2020.109476

Shailaja, G. K., & Rao, C. G. (2019). Opposition intensity-based cuckoo search algorithm for data privacy preservation. *Journal of Intelligent Systems*, *29*(1), 1441–1452. doi:10.1515/jisys-2018-0420

Sheela, M. A., & Vijayalakshmi, K. (2017). Partition based perturbation for privacy preserving distributed data mining. *Cybernetics and Information Technologies*, *17*(2), 44–55. doi:10.1515/cait-2017-0015

Skarkala, M. E., Maragoudakis, M., Gritzalis, S., & Mitrou, L. (n.d.). *PP-TAN: a Privacy Preserving Multi-party Tree Augmented Naive Bayes Classifier*. Academic Press.

Sui, P., & Li, X. (2017). A privacy-preserving approach for multimodal transaction data integrated analysis. *Neurocomputing*, *253*, 56–64. doi:10.1016/j.neucom.2016.09.130

Sun, C., Gao, H., Zhou, J., Fu, Y., & She, L. (2014). A new hybrid approach for privacy preserving distributed data mining. *IEICE Transactions on Information and Systems*, *97*(4), 876–883. doi:10.1587/transinf.E97.D.876

Tran, N. H., Le-Khac, N. A., & Kechadi, M. T. (2020). Lightweight Privacy-Preserving Data Classification. *Computers & Security*, *97*, 101835. doi:10.1016/j.cose.2020.101835

UCI. (n.d.). https://archive.ics.uci.edu/ml/datasets.php

Upadhyay, S., Sharma, C., Sharma, P., Bharadwaj, P., & Seeja, K. R. (2018). Privacy preserving data mining with 3-D rotation transformation. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 524–530. doi:10.1016/j.jksuci.2016.11.009

Wei, R., Tian, H., & Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. *Computers & Electrical Engineering*, *67*, 509–519. doi:10.1016/j.compeleceng.2018.02.017

Wu, J. M. T., Lin, C. W., Fournier-Viger, P., Djenouri, Y., Chen, C. H., & Li, Z. (2019). *The density-based clustering method for privacy-preserving data mining*. Academic Press.

Xia, Y., Zhu, T., Ding, X., Jin, H., & Zou, D. (2019). Heterogeneous differential privacy for vertically partitioned databases. *Concurrency and Computation*, e5607. doi:10.1002/cpe.5607

Yun, U., & Kim, J. (2015). A fast perturbation algorithm using tree structure for privacy preserving utility mining. *Expert Systems with Applications*, *42*(3), 1149–1165. doi:10.1016/j.eswa.2014.08.037

Yunhong, H., Liang, F., & Guoping, H. (2009, August). Privacy-preserving SVM classification on vertically partitioned data without secure multi-party computation. In *2009 fifth international conference on natural computation* (Vol. 1, pp. 543-546). IEEE. doi:10.1109/ICNC.2009.120

Zhou, J., Cao, Z., Dong, X., & Lin, X. (2015). PPDM: A privacy-preserving protocol for cloud-assisted e-healthcare systems. *IEEE Journal of Selected Topics in Signal Processing*, *9*(7), 1332–1344. doi:10.1109/JSTSP.2015.2427113

Shivlal Mewada (Member, IEEE) is working with the Department of Computer Science, Govt. Holkar (Model, Autonomous) Science College- Indore since 2011 and is a Principal founder of ISROSET, India. Mewada received Ph.D. degree in Computer Science from Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, Chitrakoot. He shared the responsibility of the research activities, coordinator of M.Phil. and is currently supervising/co-supervising M.Phil. Students at the Department of CS, Govt. Holkar Science College- Indore. He has received a prestigious Junior Research Fellow (JRF) by the Indian Government under the UGC Fellow scheme, UGC, New Delhi, India in 2011. He has two international patents in his credit; those have been granted by the Government of Australia in 2020. He has been a member of IEEE since 2013. He is also a technical committee and editorial member of various reputed international journals including Taylor and Francis, Inderscience, IEEE, and Springer conferences. He chaired 3 national seminars/conferences and 4 international conferences. He is an astute academician and has organized 2 special sessions for international conferences. He also contributed to the organization of 3 national webinars, 3 national and 5 international conferences. Dr. Mewada has published 2 book chapters in IGI Global and over 28 research articles in peer-reviewed journals like E-SCI, SCI, and Scopus including IEEE conferences. His areas of interest include; Cryptography, Information Security, Cloud Computing, IoT, and Computational Intelligence based education. He has 9 years of teaching experience and 9 years of Research Experience.