Performance Analysis of Perturbation-based Privacy Preserving Techniques: An Experimental Perspective

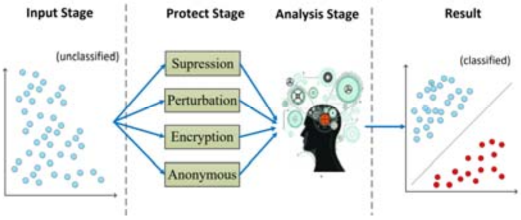
Ritu Ratra[†], Preeti Gulia[†], Nasib Singh Gill[†]
ritu.rs.dcsa@mdurohtak.ac.in preeti@mdurohtak.ac.in nasib.gill@mdurohtak.ac.in
[†] Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India

Abstract
In the present scenario, enormous amounts of data are produced every second. These data also contain private information from sources including media platforms, the banking sector, finance, healthcare, and criminal histories. Data mining is a method for looking through and analyzing massive volumes of data to find usable information. Preserving personal data during data mining has become difficult, thus privacy-preserving data mining (PPDM) is used to do so. Data perturbation is one of the several tactics used by the PPDM data privacy protection mechanism. In Perturbation, datasets are perturbed in order to preserve personal information. Both data accuracy and data privacy are addressed by it. This paper will explore and compare several perturbation strategies that may be used to protect data privacy. For this experiment, two perturbation techniques based on random projection and principal component analysis were used. These techniques include Improved Random Projection Perturbation (IRPP) and Enhanced Principal Component Analysis based Technique (EPCAT). The Naive Bayes classification algorithm is used for data mining approaches. These methods are employed to assess the precision, run time, and accuracy of the experimental results. The best perturbation method in the Nave-Bayes classification is determined to be a random projection-based technique (IRPP) for both the cardiovascular and hypothyroid datasets.

Keywords:
*Privacy-Preserving, Princ*ipal Component Analysis, Random *Projection, Differential privacy, Perturbation.*

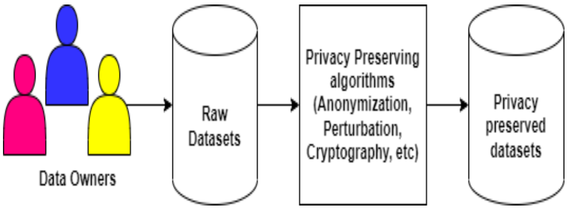
1. Introduction

Mining datasets spread across numerous parties without disclosing further private information has recently gained importance. In many businesses, protecting data privacy is currently a big concern for the data mining process. Many people are now worried that their personal information may be exposed and used for improper purposes. They think that individuals' private information should be protected [1], [2]. Additionally, safeguards for personal data protection should be in place. Privacy-preserving data mining tools have been proven and put into practice to solve this issue [3]. The importance of PPDM technology, which allows data mining without disclosing the features of the original data, is rising. Security assurance solutions that have been developed based on a number of annoyances are being combined using a variety of data mining techniques. PPDM helps to safeguard private data and sensitive information for individuals. In Fig. 1(a), the PPDM structure is shown. In order to assure information preservation, this paper aims to apply perturbation methods. To prevent the recovery of the original dataset values, data perturbation methods have changed the values of dataset record values. Additionally, it preserves the advantageous features of the dataset. The properties of the dataset are preserved using swapping, condensation, randomised response, additive noise, and other methods [4]. To preserve record-level patterns, some techniques experiment. By substituting some alternative data that are similar to those of records with comparable non-sensitive data, all sensitive data is replaced. It can be carried out using either the distributions of sensitive data when specific non-sensitive data are present the mean of sensitive data [5]. The pair-wise distances of the dataset's records are preserved through a few unique techniques, including geometric data perturbation and random projection. This means that these techniques are increasingly useful to other data mining jobs, such as classification and regression, where predictions about specific records are made [6]. These transformations and conversions typically include numerical data. Additionally, some procedures entail simple changes [7]. By transforming user input into an improbable and unpredictable form, perturbation techniques have been created as a way to ensure secrecy. Perturbation approach involves altering the original dataset's structure or introducing a little amount of noise to the data. To allow someone to receive data that has been slightly altered is known as data perturbation [8]. The data can only be modified by authorized individuals, as illustrated in the fig 1(a), 1(b). He then makes the data available to analysts for the data mining process [9].



(a)

Manuscript received October 5, 2023
Manuscript revised October 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.10.11


Data Owners
Raw Datasets
Privacy Preserving algorithms (Anonymization, Perturbation, Cryptography, etc)
Privacy preserved datasets

(b)

Fig1 (a) PPDM Structure (b) Framework of Privacy Preserving

1.1 Perturbation

Data perturbation may be utilized to efficiently employ PPDM. It is one of the often used techniques for protecting privacy [10]. In the perturbation mechanism, data is altered before processing. There are several ways of perturbation. A range of methods, including as noise addition, data hiding techniques, swapping, and many more, may be used to change the information in datasets [11]. The most effective and successful approach among these alternatives is data disruption. The perturbation methodology may be separated into two categories: the probability distribution method and the value distortion method [12], [13].

In the first technique, the data is immediately replaced by the distribution, however in the value distortion method, the data is directly altered either by using another randomization process or by introducing noise. Rotation perturbation, projection perturbation, and geometric data perturbation are the three types into which data perturbation may be divided [14]. In projection perturbation, modification is accomplished by changing the dimensions. Data randomly moves in this manner from high-dimensional to low-dimensional space [15]. In the geometric perturbation approach, perturbation is performed using a mixture of several techniques, including rotation transformation, translation transformation, and adding random value [16] - [18].

The importance of the data value is maintained despite rotation transformation and perturbation being applied to two different relational attributes. Users of real-world data may have different requirements and rights to utilize the same data collection. When publishing data, the data owner may employ a variety of privacy-preserving techniques. Data that has been mildly, moderately, or considerably changed may be shared with different types of users by the data owner. Owner can use various perturbation strengths to change the datasets [19]. This research offered a comparison of approaches based on random projection and principal component analysis that concurrently improve data classification accuracy while lowering the high dimension to a low dimension in order to safeguard the dataset's privacy [20], [21] .

The performance of two perturbation-based privacy-preserving methods is examined in the current research. Healthcare datasets have been used to test this analysis. The following are this paper's main contributions:

- The significance of privacy in data publishing are discussed in this study.
- The current study presents a thorough analysis of PPDM techniques based on perturbation.
- This paper includes the analysis of two perturbation-based privacy-preserving methods i.e. Improved Random Projection Perturbation (IRPP) and Enhanced Principal Component Analysis based Technique (EPCAT).
- This paper provides an evaluation and comparison of the perturbed dataset with the original dataset.

The rest of the paper is structured as follows: The literature review on privacy protection is explained in section 2. Improved Random Projection Perturbation (IRPP) and Enhanced Principal Component Analysis based Technique (EPCAT), two perturbation-based privacy-preserving approaches, are presented in Section 3. Section 4 provides the description of the experimental results and its comparisons with the existing work. Section 5 concludes the research paper in the end.

2. Literature Survey

Dataset privacy and security have been the subject of extensive investigation. PPDM-related methodologies and strategies have been proposed and used in the past in a variety of ways. However, the majority of these strategies don't work in all situations.

For the purpose of applying k-anonymization to the resulting structured records, Brijesh Mehta and Rao discovered existing ways from the field of Natural Language Processing (NLP) to convert the unstructured data to a structured form [22] Recognition (NER) technique is used. To anonymize the well-represented unstructured data and enable privacy-preserving unstructured big data publishing, they developed an Improved Scalable k-Anonymization (ImSKA). The results indicate that the proposed solutions beat the existing approaches in terms of F1 score and Normalized Cardinality Penalty (NCP), respectively, for both of the proposed approaches—NER and ImSKA.

A perturbation-based method for protecting privacy in data mining was presented by [23] Sangeetha Mariammala et al. in 2021. Its foundation is the additive rotation strategy. They calculated the privacy level using the variance of the initial dataset. After using their perturbation-based approach on the original dataset, it was found that its protection was improved. R. V. Banu and N. Nagaveni provided a perturbation-based method to protect privacy in data mining. Its foundation is the additive rotation strategy. They calculated the privacy level using the variance of the initial dataset. After using their perturbation-based approach on the original dataset, it was found that its protection was improved [24]. Using a hybrid approach, Vibhor Sharma et al. [25] suggested protecting privacy while data mining. Their suggested approach combines randomization and suppression. It is stated that this method recovers the original data value while maintaining data privacy and preventing information loss.

P. R. M. Rao et al. demonstrated that this strategy is more efficient and scalable. They contrasted their algorithm with the anonymization strategy and claimed that it is not vulnerable to various attacks. Additionally, several trials show that their algorithm provided 100% data usefulness [26]. A.Viji Amutha Mary [27] asserts that the Random projection strategy has a higher level of privacy than the other approaches. The photos can be very well maintained by employing RP. This method makes it possible to protect data better. It is possible to increase privacy. S. Ghosh et al. [28] gave a thorough analysis of the currently employed PPDM approaches and categorized the different data modification techniques. They used comparisons to explain the benefits and drawbacks of the various PPDM approaches. This review study touched on the PPDM's current issues, difficulties, and certain unresolved problems.

The usage of methods based on Secure Multiparty Computation is typically computationally expensive, according to M. Al-Rubaie et al., techniques that can be applied to data streams must be specifically created for particular PPDM algorithm types. The various sorts of anonymization strategies were discussed by Pervez Eager and colleagues. These methods are founded on concepts like k-anonymity, t-closeness, etc. These techniques were used during data mining, and anonymization was done when records from various sources were combined [29].
S. Mariammal [30] provided a practical hybrid method for safeguarding the dataset's privacy. For numerical data, they employ geometric data perturbation, and for categorical data, they use the k anonymization technique (generalisation). In their approach, GeethaMarya and. Iyengara performed perturbation with randomization. Intervals were used to create the data. The accuracy of the dataset was well maintained after they applied a classification algorithm to the updated data set.

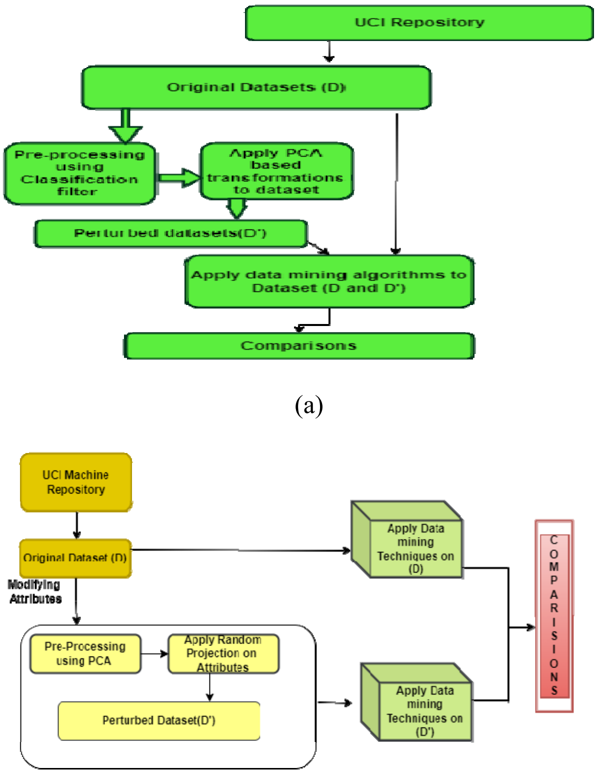
A. Pika [31] investigated a number of data perturbation techniques. In data perturbation methods, records' data values are changed. Their research indicates that the perturbation approach utilized to protect the confidentiality of original values. Dataset characteristics were preserved by techniques such rank switching, condensation, randomised response, and additive noise. Through their investigation they came to the conclusion that methods like geometric perturbation and random projection preserved the pair-wise distances between records. By doing this, they become more beneficial for data mining procedures. He further mentioned that these techniques just required basic record changes. The continual use of data sets was increasing their effectiveness.

3. Methodology

There has been a lot of research about the privacy and security of datasets. A variety of PPDM-related methodologies and procedures have already been developed and applied [32], [33]. In this article two techniques named Enhanced Principal Component Analysis based Technique [34] and Improved Random Projection Perturbation are discussed and compared.

3.1 Enhanced Principal Component Analysis based Technique (EPCAT)

It is a PCA- and classification-based approach that protects privacy. In this method, the initial stage involves pre-processing the original data using a data filter. The filtered data is then subjected to a PCA-based modification after the data pre-processing stage. Finally, the modified data is subjected to a classification approach (Naive Bayes) for data mining. The following Fig 2 (a) shows the functional flow diagram for this model.


UCI Repository
Original Datasets (D)
Pre-processing using Classification filter
Apply PCA based transformations to dataset
Perturbed datasets(D')
Apply data mining algorithms to Dataset (D and D')
Comparisons
UCI Machine Repository
Original Dataset (D)
Modifying Attributes
Pre-Processing using PCA
Apply Random Projection on Attributes
Perturbed Dataset(D')
Apply Data mining Techniques on (D)
Apply Data mining Techniques on (D')
COMPARISIONS

(a)

(b)

Fig2: (a) Framework of EPCAT, (b) Framework of IRPP

The following two phases make up the full structure of this technique:

Phase 1: The preservation of individual privacy in datasets is the focus of this phase. This phase consists of mainly two parts. Which are:

a) The most crucial component for improving the precision and speed of the classification approach is the classification filter module (or CFM). Prior to the PCA modifications of the dataset, this filter is applied to the original dataset.

b) The second module is the perturbation module, where the altered data set is once more disturbed using PCA-based transformations. Additionally examined and contrasted with the original dataset is the affected dataset's correctness.

Phase 2: Classification process is performed on affected datasets at this phase.

Module for classification: The perturbed data set is mined after the two aforementioned modules. The "Naive Bayes" approach is used as the classification method in this instance. Additionally, accuracy is calculated on the original datasets and contrasted with the accuracy of the perturbed dataset..

3.2 Improved Random Projection Perturbation (IRPP)

Random projection is a potent method for reducing dimensionality of dataset. Random projection involves utilizing a random k × d matrix to project the original high d-dimensional data onto a smaller k-dimensional subspace. Fig 2(b) provides a general overview of the Improved Random Projection Perturbation method. The paper's overall structure is split into two sections.

Phase 1: The preservation of individual privacy in datasets is the focus of this phase. Essentially, this phase consists of two parts. Which are:

a) Feature selection: This module is used to choose features and improve the classification technique's accuracy. Prior to the dataset's changes using Random Projection, this was done to the original dataset using Principal Component Analysis (PCA). Prior to the Random Projection process and the classification phase, feature selection is used. In this paper, a feature selection method based on PCA is applied for seletion of appropriate features.

b) Random Projection: The perturbed data is adjusted once more in this module utilizing dimension reduction, which is accomplished with the aid of the random projection method. The datasets are distorted using the random projection technique. Additionally examined and contrasted with the original dataset is the affected dataset's correctness.

Phase 2: In this phase, the classification process is applied to altered datasets.

Module for classification: After the two aforementioned processes have been implemented, perturbed data sets are mined using a particular classification technique. As classification algorithms, NaiveBased is used in this instance. Additionally, several matrices are computed on the original datasets and their accuracy is contrasted with that of the perturbed dataset. The algorithm IPRR technique is described in fig 3 given below:

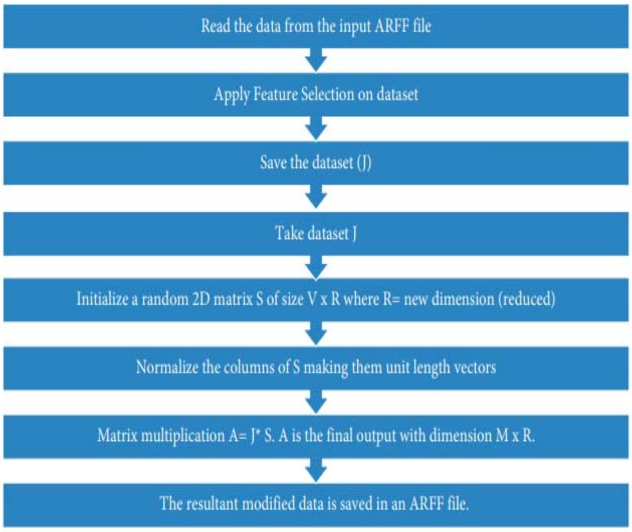

Read the data from the input ARFF file
Apply Feature Selection on dataset
Save the dataset (J)
Take dataset J
Initialize a random 2D matrix S of size V x R where R= new dimension (reduced)
Normalize the columns of S making them unit length vectors
Matrix multiplication A= J* S. A is the final output with dimension M x R.
The resultant modified data is saved in an ARFF file.

Fig3: Flowchart of IRPP Technique

4. Results and Discussion

By using WEKA, the performance analysis is carried out [35]. The original datasets are used in these experimental analyses together with the chosen methods to create the transformed datasets. Dataset classification with machine learning is a highly beneficial process. Machine learning offers a variety of classification techniques, including logistic regression, naive bayes, decision trees, etc. [36]. The Naive Bayes classification method is used for implementations of the algorithms in order to evaluate the efficacy and efficiency of the strategies. On both datasets, numerous metrics including accuracy, model building time, kappa static measure, mean absolute error, and f-measure are calculated using different techniques. These metrics are used to assess how well the chosen algorithms performed on the projected dataset.

Datasets: In this paper, two datasets are used for experimental purposes. These datasets are the cardiovascular disease dataset [37] and the hypothyroid disease dataset. The description of both dataset is provided in Table 1.

Table 1: Description of Datasets

Name of Dataset	***No. of Instances***	***No. of Attributes***	***Attribute Description***
Cardiovascular disease dataset	70k	13	Id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco , active, class
Hypothyroid disease dataset	7200	21	Age, sex, on thyroxine, query on thyroxine, on antithyroid medication, sick, pregnant, thyroid surgery, I131 treatment, Query hypothyroid, query hyperthyroid, lithium, goiter, tumor, hypopituitary, psych TSH measured, TSH, T3 measured, T3, TT4 measured, TT4, T4U measured, T4U, FTI measured, Class

This research compares two perturbation-based privacy-preserving approaches. Several tests were run on data sets of two different sizes, and the associated outcomes were seen. The results of the experiments demonstrate that the Improved Random Projection Perturbation (IRPP) approach performs better due to its greater accuracy, TP rate, FP rate, F-Measures, and run duration values. The effectiveness of the suggested method to the conventional classification model on cardiovascular datasets is shown in Table 2. On the provided training datasets, the metrics accuracy, TP rate, FP rate, F-Measures, and run time are computed. It is clearly shown in the table and can be seen that the suggested strategy produces superior results than the conventional model of categorization in all regards.

Table 2: Performance measure of classification algorithms on Cardiovascular dataset

Accuracy measurement	***Naïve Bayes Classifier(on Original dataset)***	***Naïve Bayes Classifier(on IRPP based Perturbed dataset)***	***Naïve Bayes Classifier(on EPCAT based Perturbed dataset)***
Time taken to build (sec)	0.53	0.34	0.25
Correctly Classified Instances (%)	58.85	59.87	59.65
Incorrectly Classified Instances (%)	41.15	40.13	40.31
TP Rate (%)	58.90	59.70	59.16
F-Measures (%)	54.40	56.50	56.14

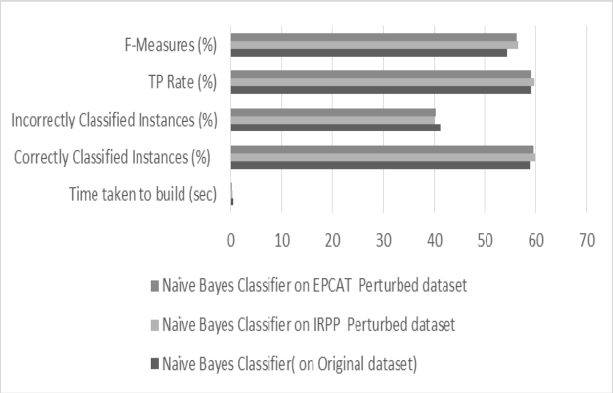


Fig 4: Performance Analysis IRPP and EPCAT to the Conventional Model on Cardiovascular dataset

The performance of IRPP and EPCAT privacy-preserving algorithms to the conventional classification algorithms on Cardiovascular datasets is shown in Fig 4. It displays the efficiency of the random projection-based privacy-preserving and PCA-based privacy-preserving method to the traditional classification model on Cardiovascular datasets. As shown in the figures, it is observed that the IRPP method has better accuracy measures than the conventional classification algorithms and PCA-based privacy-preserving method.

The effectiveness of both methods to the conventional classification model on hypothyriod dataset is shown in Table 3. On the provided training datasets, the metrics accuracy, TP rate, FP rate, F-Measures, and run time are calculated. It is well depicted in the in the table, and it is easy to see that the IRPP approach yields better results overall than the conventional model of classification.

Table 3: Performance measure of classification on Hypothyroid dataset

Accuracy measurement	***Naïve Bayes Classifier(on Original dataset)***	***Naïve Bayes Classifier(on IRPP based Perturbed dataset)***	***Naïve Bayes Classifier(on EPCAT based Perturbed dataset)***
Correctly Classified Instances (%)	71.67	75.67	74.08
Incorrectly Classified Instances (%)	28.32	24.32	25.91
TP Rate (%)	71.70	75.13	74.10
F-Measures (%)	70.80	73.54	73.48

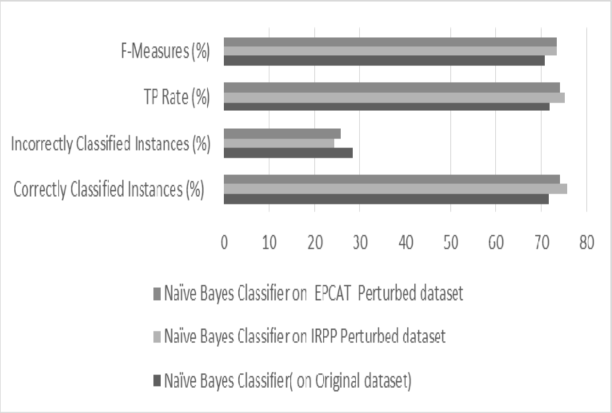


Fig5: Performance Analysis IRPP and EPCAT to the Conventional Model on Hypothyroid dataset using Naïve Bayes classifier

On hypothyroid datasets, Fig 5 compares the efficiency of the privacy-preserving algorithms IRPP and EPCAT to the traditional classification techniques. It demonstrates the effectiveness of the privacy-preserving random projection and PCA methods in comparison to the conventional classification model on hypothyroid datasets. As seen in the fig, it can be seen that the IRPP approach outperforms both the PCA-based privacy-preserving method and traditional classification algorithms in terms of accuracy measurements.

5. Conclusion

The main purpose of privacy preserving in data mining processes is to develop algorithms that can hide or provide privacy to some sensitive information to prevent unauthorized access by profiteers. However, privacy and accuracy in data mining conflict. In this regard, this paper has compared the number of PPDM techniques based on perturbation. This article provides a brief overview of some privacy techniques, namely PCA-based perturbation, and random projection-based perturbation, and analyzes their competencies and differences in different scenarios. On different large datasets, the usefulness and accuracy of both techniques have been tested in classification algorithms Naive Bayed classifiers. With the use of diverse experimental results, it has been determined that IRPP (Improved Random Projection Perturbation) privacy-preserving technique is more accurate and efficient than EPCAT (Enhanced Principal Component Analysis based Technique) and traditional techniques. In the case of cardiovascular datasets, this technique outperforms EPCAT and traditional techniques in terms of runtime, accuracy, TP rate, and F-measurer. In the case of the hypothyroid dataset, experimental outcomes on all measurements (efficiency, accuracy, run time, TP rate, and F-measurer) are better or almost identical to the previous approach model.

Acknowledgment

The authors are thankful to the Department of Computer science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India for its support. The authors are grateful to https://archive.ics.uci.edu/ml/index.php and https://www.kaggle.com/datasets for providing the datasets for experimental purposes.

References

[1] A. Altalhi, M. AL-Saedi, H. Alsuwat, and E. Alsuwat, “Privacy-Preserving in the Context of Data Mining and Deep Learning,” Int. J. Comput. Sci. Netw. Secur., vol. 21, no. 6, pp. 137–142, 2021.

[2] M. Thottipalayam Andavan and N. Vairaperumal, “Privacy protection domain-user integra tag deduplication in cloud data server,” Int. J. Electr. Comput. Eng. IJECE, vol. 12, no. 4, p. 4155, Aug. 2022, doi: 10.11591/ijece.v12i4.pp4155-4163.

[3] P. Gulia, “Privacy Preserving Data Mining Of Vertically Partitioned Data In Distributed Environment-An Experimental Analysis,” J. Theor. Appl. Inf. Technol., vol. 96, no. 10, 2018.

[4] R. Ratra and P. Gulia, “Privacy Preserving Data Mining: Techniques and Algorithms,” Int. J. Eng. Trends Technol., vol. 68, no. 11, pp. 56–62, Nov. 2020, doi: 10.14445/22315381/IJETT-V68I11P207.

[5] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “Random-data perturbation techniques and privacy-preserving data mining,” Knowl. Inf. Syst., vol. 7, no. 4, pp. 387–414, 2005.

[6] K. Liu, H. Kargupta, and J. Ryan, “Random projection-based multiplicative data perturbation for privacy preserving distributed data mining,” IEEE Trans. Knowl. Data Eng., vol. 18, no. 1, pp. 92–106, 2005.

[7] N. Gayathri Devi and K. Manikandan, “Improved perturbation technique privacy-preserving rotation-based condensation algorithm for privacy preserving in big data stream using Internet of Things,” Trans. Emerg. Telecommun. Technol., vol. 31, no. 12, pp. 1–12, 2020.

[8] K. M. Chong, “Privacy-preserving healthcare informatics: a review,” ITM Web Conf., vol. 36, p. 04005, 2021, doi: 10.1051/itmconf/20213604005.

[9] M. Al-Rubaie, P. Wu, J. M. Chang, and S.-Y. Kung, “Privacy-preserving PCA on horizontally-partitioned data,” in 2017 IEEE Conference on Dependable and Secure Computing, 2017, pp. 280–287.

[10] M. Dabhade and J. J. Hilda, “Privacy Preserving In Data Mining Using Data Perturbation And Classification Method,” Ecer Iioabj, vol. 8, pp. 346–352.

[11] C. Eyupoglu, M. A. Aydin, A. H. Zaim, and A. Sertbas, “An efficient big data anonymization algorithm based on chaos and perturbation techniques,” Entropy, vol. 20, no. 5, p. 373, 2018.

[12] D. El Majdoubi, H. El Bakkali, S. Sadki, Z. Maqour, and A. Leghmid, “The Systematic Literature Review of Privacy-Preserving Solutions in Smart Healthcare

Environment," Secur. Commun. Netw., vol. 2022, pp. 1–26, Mar. 2022, doi: 10.1155/2022/5642026.

[13] O. Mir, M. Roland, and R. Mayrhofer, "Decentralized, Privacy-Preserving, Single Sign-On," Secur. Commun. Netw., vol. 2022, pp. 1–18, Jan. 2022, doi: 10.1155/2022/9983995.

[14] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: current scenario and future prospects," in 2012 third international conference on computer and communication technology, 2012, pp. 26–32.

[15] G. D. N and M. K, ″Improved perturbation technique privacy - preserving rotation - based condensation algorithm for privacy preserving in big data stream using Internet of Things,″ Trans. Emerg. Telecommun. Technol., vol. 31, no. 12, Dec. 2020, doi: 10.1002/ett.3970.

[16] R. Ratra, P. Gulia, and N. S. Gill, "Evaluation of Re-identification Risk using Anonymization and Differential Privacy in Healthcare," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 2, 2022.

[17] W. Shen, Q. Guo, H. Zhu, K. Tang, S. Zhan, and Z. Hao, "The Privacy Data Protection Model Based on Random Projection Technology," in Big Data and Security, Singapore, 2021, pp. 215–226. doi: 10.1007/978-981-16-3150-4_19.

[18] X. Fan, G. Wang, K. Chen, X. He, and W. Xu, "Ppca: Privacy-preserving principal component analysis using secure multiparty computation (mpc)," ArXiv Prepr. ArXiv210507612, 2021.

[19] C. Sun, L. Ippel, A. Dekker, M. Dumontier, and J. van Soest, "A systematic review on privacy-preserving distributed data mining," Data Sci., no. Preprint, pp. 1–30.

[20] V. Sharma, D. Soni, D. Srivastava, and P. Kumar, "A Novel Hybrid Approach of Suppression and Randomization for Privacy Preserving Data Mining.," Ilk. Online, vol. 20, no. 5, 2021.

[21] P. H. Li, T. Lee, and H. Y. Youn, "Dimensionality Reduction with Sparse Locality for Principal Component Analysis," Math. Probl. Eng., vol. 2020, pp. 1–12, May 2020, doi: 10.1155/2020/9723279.

[22] B. B. Mehta and U. P. Rao, "Improved l-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1423–1430, Apr. 2022, doi: 10.1016/j.jksuci.2019.08.006.

[23] S. Mariammal, "An Additive Rotational Perturbation Technique for Privacy Preserving Data Mining," *Turk. J. Comput. Math. Educ. TURCOMAT*, vol. 12, no. 9, pp. 2675–2681, 2021.

[24] R. V. Banu and N. Nagaveni, "Preservation of data privacy using PCA based transformation," in 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009, pp. 439–443..

[25] V. Sharma, D. Soni, D. Srivastava, and P. Kumar, "A Novel Hybrid Approach of Suppression and Randomization for Privacy Preserving Data Mining." Ilk. Online, vol. 20, no. 5, 2021.

[26] Department of Computer Science and Engineering, JNTUA, Anantapuramu, Andhra Pradesh, India, P. R. M. Rao, S. M. Krishna, and A. P. S. Kumar, "Novel algorithm for efficient privacy preservation in data analytics," Indian J. Sci. Technol., vol. 14, no. 6, pp. 519–526, Feb. 2021, doi: 10.17485/IJST/v14i6.1773.

[27] Associate Professor, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai-600119, India. and Av. Mary, "A Random Projection Approach To Secure Medical Images.," Int. J. Adv. Res., vol. 7, no. 3, pp. 1298–1301, Mar. 2019, doi: 10.21474/IJAR01/8763.

[28] S. Ghosh, S. Sadhu, S. Biswas, D. Sarkar, and P. P. Sarkar, "A Comparison between Different Classifiers for Tennis Match Result Prediction," Malays. J. Comput. Sci., vol. 32, no. 2, pp. 97–111, Apr. 2019, doi: 10.22452/mjcs.vol32no2.2.

[29] M. Al-Rubaie, P. Wu, J. M. Chang, and S.-Y. Kung, "Privacy-preserving PCA on horizontally-partitioned data," in 2017 IEEE Conference on Dependable and Secure Computing, 2017, pp. 280–287.

[30] S. Mariammal, "An Additive Rotational Perturbation Technique for Privacy Preserving Data Mining," Turk. J. Comput. Math. Educ. TURCOMAT, vol. 12, no. 9, pp. 2675–2681, 2021.

[31] A. Pika, M. T. Wynn, S. Budiono, A. H. M. ter Hofstede, W. M. P. van der Aalst, and H. A. Reijers, "Privacy-Preserving Process Mining in Healthcare," Int. J. Environ. Res. Public. Health, vol. 17, no. 5, p. 1612, Mar. 2020, doi: 10.3390/ijerph17051612.

[32] S. Patel, G. Shah, and A. Patel, "'Techniques of data perturbation for privacy preserving data mining," Int J Advent Res Comput Electron, vol. 1, pp. 5–10, 2014

[33] X. Liu, Y. Lin, Q. Liu, and X. Yao, "A Privacy-Preserving Principal Component Analysis Outsourcing Framework," in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, USA, Aug. 2018, pp. 1354–1359. doi: 10.1109/TrustCom/BigDataSE.2018.00187.

[34] R. Ratra, P. Gulia, N. S. Gill, and J. M. Chatterjee, "Big Data Privacy Preservation Using Principal Component Analysis and Random Projection in Healthcare," Math. Probl. Eng., vol. 2022, p. 6402274, Aug. 2022, doi: 10.1155/2022/6402274.

[35] R. Ratra and P. Gulia, "Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange)," Int. J. Eng. Trends Technol., vol. 68, no. 8, pp. 30–35, Aug. 2020, doi: 10.14445/22315381/IJETT-V68I8P206S.

[36] A. Amkor, K. Maaider, and N. El Barbri, "An evaluation of machine learning algorithms coupled to an electronic olfactory system: a study of the mint case," Int. J. Electr. Comput. Eng. IJECE, vol. 12, no. 4, p. 4335, Aug. 2022, doi: 10.11591/ijece.v12i4.pp4335-4344.

[37] "Find Open Datasets and Machine Learning Projects | Kaggle." https://www.kaggle.com/datasets (accessed May 31, 2022).

[38] S. K. David, M. Rafiullah, and K. Siddiqui, "Comparison of Different Machine Learning Techniques to Predict Diabetic Kidney Disease," J. Healthc. Eng., vol. 2022, pp. 1–9, Apr. 2022, doi: 10.1155/2022/7378307

Ritu Ratra received the MCA degrees from Maharshi Dayanand University, Rohtak, Haryana, India. Currently, she is pursuing her Ph.D. in Computer Science at the Department of Computer Science & Applications, Maharshi Dayanand University. Rohtak, Haryana, India. Previously, she had worked as an Assistance Professor at Sh. Lal Nath Hindu College, Rohtak, Haryana, India affiliated with MD University, Rohtak, and Haryana, India for 12.5 years. Her research areas include machine learning, Data mining, Big Data, and Artificial Intelligent. She has authored or coauthored many refereed journal and conference papers, she can be contacted at email: ritu.rs.dcsa@mdurohtak.ac.in.ORCID ID: 0000-0001-8018-5460.

Preeti Gulia received Ph.D. degree in computer science in 2013. She is currently working as Associate Professor at the Department of Computer Science & Applications, M.D. University, Rohtak, Haryana, India. She is serving the Department since 2009. She has published more than 65 research papers and articles in journals and conferences of National/ International repute including ACM, and Scopus. Her area of research includes Data Mining, Big Data, Machine Learning, Deep Learning, IoT, and Software Engineering. She is an active professional member of IAENG, CSI, and ACM. She is also serving as Editorial Board Member Active Reviewer of International/ National Journals. She has guided one research scholar as well as guiding four Ph.D. research scholars from various research areas. She can be contacted at email: preeti@mdurohtak.ac.in , ORCID ID: 0000-0001-8535-4016.

Nasib S. Gill holds Post-Doctoral research in Computer Science at Brunel University, West London during 2001-2002 and Ph.D. in Computer Science in 1996. He is a recipient of the Commonwealth Fellowship Award of the British Government for the Year 2001. Besides, he also has earned his MBA degree. He is currently Head, Department of Computer Science & Applications, M. D. University, Rohtak, India. He is also working as Director, Directorate of Distance Education as well as Director of Digital Learning Centre, M. D. University, Rohtak, Haryana. He is an active professional member of IETE, IAENG, and CSI. He has published more than 304 research papers and authored 5 popular books He has guided so far 12 Ph.D. scholars as well as guiding about 5 more scholars. His research interests primarily include – IoT, Machine & Deep Learning, Information and Network Security, Data mining & Data warehousing, NLP, Measurement of Component-based Systems, etc. He can be contacted at email: nasib.gill@mdurohtak.ac.in , ORCID ID: 0000-0002-8594-4320.