Performance Analysis of Proposed Scalable Reversible Randomization Algorithm (SRRA) in Privacy Preserving Big Data Analytics

Mohana Chelvan P[1], Dr. Rajavarman V N[2], Dr. Dahlia Sam[3]
Research Scholar, Department of Information Technology, Faculty of Engineering and Technology, Dr. M.G.R. Educational and Research Institute, Chennai, India[1]
Senior Assistant Professor, School of Science and Computer Studies, CMR University, Bengaluru, India[1]
Professor, Department of Information Technology, Faculty of Engineering and Technology, Dr. M.G.R. Educational and Research Institute, Chennai, India[2, 3]

***Abstract*—The economy of today's world is a data-driven knowledge economy, as electronic devices are mostly used for our day-to-day activities, through which organizations collect data actively or passively. The dimensionality of the dataset is also increased, along with the volume of data, because of the advancements in digital devices and communication technology. The feature selection becomes a crucial preprocessing step in big data analytics as a dimensionality reduction technique to eliminate redundant and noisy features. Studying the fluctuations in feature selection results is a vigorous area of research, as it is positively related to data utility, as fluctuations in feature selection results confuse the data analysts' minds about their research outcomes. Privacy preservation is a major concern in big data analytics to protect sensitive individuals' data. Application of privacy preservation techniques to modify the dataset will affect the stability of feature selection, as it has recently been proven that it mostly depends on the dataset's physical characteristics. This study analyses the performance of the proposed Scalable Reversible Randomization Algorithm (SRRA) in terms of privacy preservation, change in characteristics of the dataset, information loss, stability of feature selection, and data utility in big data scenarios.**

***Keywords*—Big data; data analytics; high dimensionality; feature selection; selection stability; privacy preservation; information loss**

I. INTRODUCTION

Big data is a keyword for organizations, as everyday mountains of customer data (including social media, education, healthcare, finance, business, and more) accumulate in organizations that are mostly unstructured and semi-structured. Because of the technological advancements and outsourcing of cloud computing, the dimensionality, i.e., the number of features of the dataset, is also increasing along with volume, i.e., samples. Organizations get an edge over their competitors by analyzing the data for mining patterns and getting insights, and applying the knowledge for business strategic decision-making and improved customer relationship management. However, feature selection is a vigorous preprocessing step of dimensionality reduction as it eliminates the "curse of high dimensionality" by selecting the most relevant subset of features for better comprehension of the model, higher efficiency in terms of processing time, improved accuracy of data analytics results, elimination of possible redundant and noisy attributes which may be quasi identifier attributes leading to better privacy preservation than the full set of features.

Unstable feature selection results confuse the researcher, as feature selection is related to data utility. Previously, data scientists believed that fluctuations in feature selection results depend on feature selection algorithms. However, lately, it has been proven that fluctuations in feature selection results depend mostly on changes in the physical characteristics of the dataset but are not completely independent of algorithms [1-6]. Feature selection instability is caused by minor changes in the dataset, which may select different subsets of features and may affect the accuracy of data analytics results, leading to doubts in the minds of researchers about their conclusions of research findings. Sensitive data like monetary data or disease, or marital status of individuals should be prevented from disclosure to third parties in data analytics. The application of privacy-preservation techniques affects the physical characteristics of the dataset, which will be reflected in fluctuations in feature selection results, which will be related to the accuracy of big data analytics results. Hence, there is a need to develop privacy preserving big data analytics algorithms with minimal change in the physical characteristics of the dataset for better feature selection stability results, lesser information loss, and higher data utility without compromising on privacy preservation.

All the privacy conserving methods like slicing, t-closeness, l-diversity, and k-anonymity have been practiced for a long time, but in the big data era, their applications have practical complications resulting in processing overheads because of the enormous volume, several varieties, and continuous velocity of arriving data. Traditional security mechanisms are inadequate in handling data analytics in big data and cloud computing scenarios. The tools for big data, like Hadoop, MapReduce, Pig, and Spark, work on distributed parallel processing in the Hadoop Distributed File System, leading to complications in employing privacy-conserving methods in terms of scalability and implementation overheads. It is practically very complex to implement and time-consuming in terms of processing efficiency, and it is very difficult to implement these privacy conserving methods in terms of scalability, as the data and processing algorithm are distributed in thousands of data nodes,

and also results in a diminution of the accuracy of outcomes of big data analytics, i.e., data utility.

Information security and privacy protection are very challenging in the big data and cloud computing era. To address the issues, the proposed algorithm, SRRA, is based on the randomization technique, which can be applied to the preprocessing stage itself; hence, it is highly scalable and efficient in terms of processing time for big data privacy preservation. At the preprocessing stage, the data are collected from different sources and go through data extraction, data cleansing, transformation procedures, and anonymization of sensitive and quasi-identifiers by SRRA to annihilate personal identifying information and safeguard private sensitive data, followed by feature selection as a dimensionality reduction technique.

In this research study, Section II provides the related work, and Section III describes the proposed privacy preserving big data analytics algorithm, the Scalable Reversible Randomization algorithm (SRRA). Section IV presents the methodology, results and discussion. Finally, Section V provides the conclusion.

II. Related Work

Big data analytics is vital for organizations in today's data-driven economy. However, sensitive data of individuals' records shouldn't be re-identified by the data analyst, as it can be misused. Privacy is a data owner-oriented concept and a privacy preservation method in which the dataset is under the control of a custodian called PPDM (Privacy Preserving Data Mining), and the data owner is responsible for the privacy of the individuals' data. The second one is the anonymization method, and also in this method, the dataset is not under the control of a custodian called PPDP (Privacy Preserving Data Publishing).

In PPDM, like differential privacy techniques is interactive, the dataset is under the control of the data owner, and the big data analyst analyses the dataset. However, these techniques are not appropriate for the big data era and are practically very difficult to implement because of the three v's, i.e., velocity, variety, and volume of big data, and also the data is collected from different sources. The PPDP methods like generalization, bucketization, perturbation, and anonymization techniques like k-anonymity (quasi identifiers are generalized to at least k number of alike attribute values for different records in each equivalence class), l-diversity (improved model of the k-anonymity technique, the sensitive attributes are at least l distinct domain feature values for each equivalence class that includes both generalized k number quasi identifier values and l distinct sensitive feature values to avoid both background knowledge and similarity attack), t-closeness (to eliminate skewness attack on l-diversity model), slicing, and randomization methods are non-interactive, the dataset is anonymized and published for any research works as it is not under the control of a custodian. However, these techniques are practically very difficult to implement in the distributed processing environment of big data and the cloud computing era of Hadoop, Spark, or Pig, as they are not directly applicable and also not very scalable.

The differential privacy method [7] is an interactive query made to a database using a randomized response mechanism by adding calculated noise to the outcome of the query. This method shares data to permit inferences about groups of persons while preventing someone from learning information about an individual through a rigorous mathematical foundation. It gives strong security for sensitive data even if the adversary has robust background knowledge, but less data utility because of added noise. Because of the three v's: variety, velocity, and volume of big data, the differential privacy technique is not appropriate for cloud computing and big data environments. In the Hadoop Distributed File System environment of big data, along with the distributed processing environment of the MapReduce algorithm, it is very complex and difficult to implement differential privacy in big data.

The important PPDP methods are t-closeness, l-diversity, k-anonymity, and slicing. The k-anonymity method anonymizes the data to eliminate re-identification, as every quasi-identifier is anonymized with a k number of feature values by the generalization technique [8, 9]. In the k-anonymity algorithm, the value of k of 23 means ensuring 23 blurry tuples when an effort is made to recognize a specific individual's private sensitive data. It is difficult to prevent re-identification of individuals' records by quasi-identifiers, and so disclosure of sensitive data cannot be prevented in k-anonymity. In the k-anonymity technique, record linkage attacks like homogeneity attacks and background knowledge attacks compromise the de-identification of an individual's record if there is insufficient diversity of sensitive attributes.

The l-diversity method was introduced to eliminate the attacks that compromise the k-anonymity model. There must be l well-represented sensitive feature values like disease, marital status, and monetary values in each equivalence class, in the l-diversity method [10]. Because of the different varieties and huge volume of data in big data, l-diversity is not possible at all times, and implementing l-diversity in big data is very complex and difficult. The feature disclosure cannot be safeguarded when the global distribution of data in the dataset is tilted into a few equivalence classes. Nevertheless, de-identification is compromised by the skewness attack in the l-diversity model, as in the equivalence class, one value emerges as frequently leading to the possibility for the opponent to deduce the value of another entity having a similar value with the information of real time global dispersal of data.

The t-closeness model is introduced to avoid the skewness attack in the l-diversity model. If the distance between the distributions of sensitive feature values in the class is no more than a threshold of the real skewness of the distributed data, the equivalence class is measured to have 't-closeness' in the t-closeness model [11]. These techniques have limitations and practical difficulties in big data and cloud computing scenarios. Applying these techniques is very difficult and complex in big data and clouds because of the huge volume and multiple varieties of data, and also inefficient due to long processing time, resulting in being very difficult to implement in big data frameworks like Hadoop, MapReduce, Pig, and Spark.

Slicing is the model of splitting the dataset column-wise or row-wise, i.e., splitting feature values and/or tuples where

extremely correlated features are a group and dissimilar features are sliced into different groups [12]. Slicing models are more appropriate for big data datasets with high dimensionality. Nevertheless, it is very complex to implement and also has practical limitations.

The randomization method adds calculated noise to the particular feature values [13]. Randomization methods have disadvantages in big data due to processing time efficiency and reduction in data utility, i.e., the accuracy of outcomes of big data analytics. However, the randomization method can be applied in the preprocessing step itself, and then it eliminates anonymization overhead.

Cryptographic methods such as secure multiparty computation (MPC) and homomorphic encryption are used to encrypt the data before the transfer of it for big data analytics [14]. In this method of secure multiparty computing, multiple parties can analyze the dataset together without revealing their data. It is very complex to implement cryptographic methods in big data. If multiple parties are involved in integrated big data analytics, managing the complexity is more challenging in big data. In this technique, data utility is considerably reduced by the application of encryption. Also, it is not easy in the big data and cloud computing scenario to encrypt all data.

The recent developments in privacy-conserving methods in the big data era are MapReduce-based anonymization (MRA) [15]. In this method, the dataset is considered as a single equivalence class, and in each iteration, tuples are selected for subclasses for k-anonymization. A global file is needed among the data nodes to share the updated equivalence class information with all nodes is a major drawback of the method. In the updated version of MRA, instead of a single global file, a chunk of files is used to update the equivalence class information. In this method, the drawbacks are multiple files to share equivalence class information and multiple iterations. The Scalable k-anonymization (SKA) model for MapReduce in [16]. This method has better processing time in comparison with similar existing scalable algorithms. In this method initial file is converted into a sorted equivalence class. The SKA algorithm results in considerable information loss. In the improved SKA, the columns are rearranged so that the distance between consecutive records is less, resulting in minimum information loss.

The privacy preserving big data analytics algorithm in [17] uses cryptographic techniques like homomorphic encryption schemes, attribute-based encryption schemes, and order-preserving encryption schemes in untrusted cloud environments. It is very difficult to implement this method in big data and cloud computing environments, as applying cryptographic techniques is a huge implementation overhead. This method has improved the preservation of privacy compared to other methods. The application of encryption algorithms for privacy preservation drastically reduces the utility of data.

Privacy preserving big data analytics algorithm in [18] uses two phases Named Entity Recognition (NER) approach called Conditional Random Field (CRF) to annotate unstructured data into the structured form, along with ImpSKA, an extension of the Scalable k-Anonymity (SKA) model using the Apache Pig framework. This method can be used for semi-structured and unstructured data. Practically, it is complex to implement in a big data distributed processing environment, leading to huge implementation overhead.

The improved scalable l-diversity approach in [19] is the scalable anonymization approach for privacy preserving big data publishing, and Apache Pig is used for the implementation. In this method, the input data are sorted for the creation initial equivalence class. In improved scalable l-diversity (ImSLD), both quasi-identifier attributes and sensitive attributes are used to generate an equivalence class from the sorted columns, and then the table is first k-anonymized and then l-diversified. In ImSLD, along with k-anonymization, l-diversity is also used to improve the preservation of privacy. There will be information loss and decreased data utility in comparison to the previous anonymization methods. It is a complex and time-consuming model in a big data scenario, as the required columns are arranged in ascending order of several unique values. This model also has a marginally higher processing time in comparison to similar models.

Along with generalization, anonymization (l-diversity and k-anonymity), slicing, and a method analogous to a one-way hash function called privacy view generation are used in the proposed system in [20] to safeguard the sensitive data and to avoid privacy incursions. In feature partitioning using the slicing method, extremely interrelated features are sliced as a subset of records. An equivalence class is formed for each slice of records and is anonymized by k-anonymity. Quasi identifiers are susceptible to background knowledge attacks, and hence sensitive feature values are l-diversified. However, this model is very complex to implement as it is very time-consuming in big data scenarios, leading to increased processing time and data utility is also drastically reduced.

The privacy-conserving big data analytics methods are not effective in terms of scalability, and applying these methods in big data distributed frameworks like Hadoop, MapReduce, Pig, or Spark is very intricate and time-consuming, and data utility is considerably reduced. The proposed Scalable Reversible Randomization Algorithm (SRRA) in this research study is suitable for the distributed parallel processing of big data frameworks, as it can be applied in the preprocessing stage itself, and so it is extremely scalable. It fills the gap, as the existing privacy preserving techniques are unsuitable in a big data scenario. The SRRA is practically not very difficult to implement in big data compared to the similar methods and with better fluctuations in feature selection results, lesser information loss, and higher data utility without compromising privacy preservation.

III. Proposed Scalable Reversible Randomization Algorithm (SRRA)

There will be four types of features in the dataset: non-sensitive features, indirect or quasi-identifiers, sensitive features, and direct identifiers.

Non-sensitive features are the features other than the above three types that are not sought by the intruder.

Indirect identifiers or quasi-identifiers are features like date of birth, occupation, religion, location, and zip code that can be

indirectly used for the re-identification of a record of an individual using linkage attacks.

Sensitive features are features like disease or monetary values targeted by the intruder. Privacy preservation techniques should be applied to quasi-identifiers and sensitive features to safeguard the privacy of individuals' records.

Direct Identifier attributes uniquely identify the record, like employee ID or social security number. Identifier attributes are generally anonymized before being sent for big data analytics.

The proposed reversible privacy conserving algorithm, i.e., Scalable Reversible Randomization Algorithm (SRRA), is shown in Fig. 1. SRRA is explained in [21]. Big data analytics is challenging for computational resources in terms of storing and processing, and also for machine learning algorithms in terms of efficiency and performance. Also, it is difficult to identify quasi-identifier and sensitive data in big data. The efficiency, i.e., performance of algorithms, and data utility, i.e., accuracy of results of the analysis, will be closer or the same because, with sampled versions, for the reduced samples selected by attribute selection optimization algorithms such as the ant-colony, Pareto-front, particle swarm optimization, and similar other algorithms. However, we should not be able to outperform the model of the 100% data. The privacy-conserving alteration by the proposed algorithm SRRA is scalable and efficient, along with better stability of feature selection and accuracy results.

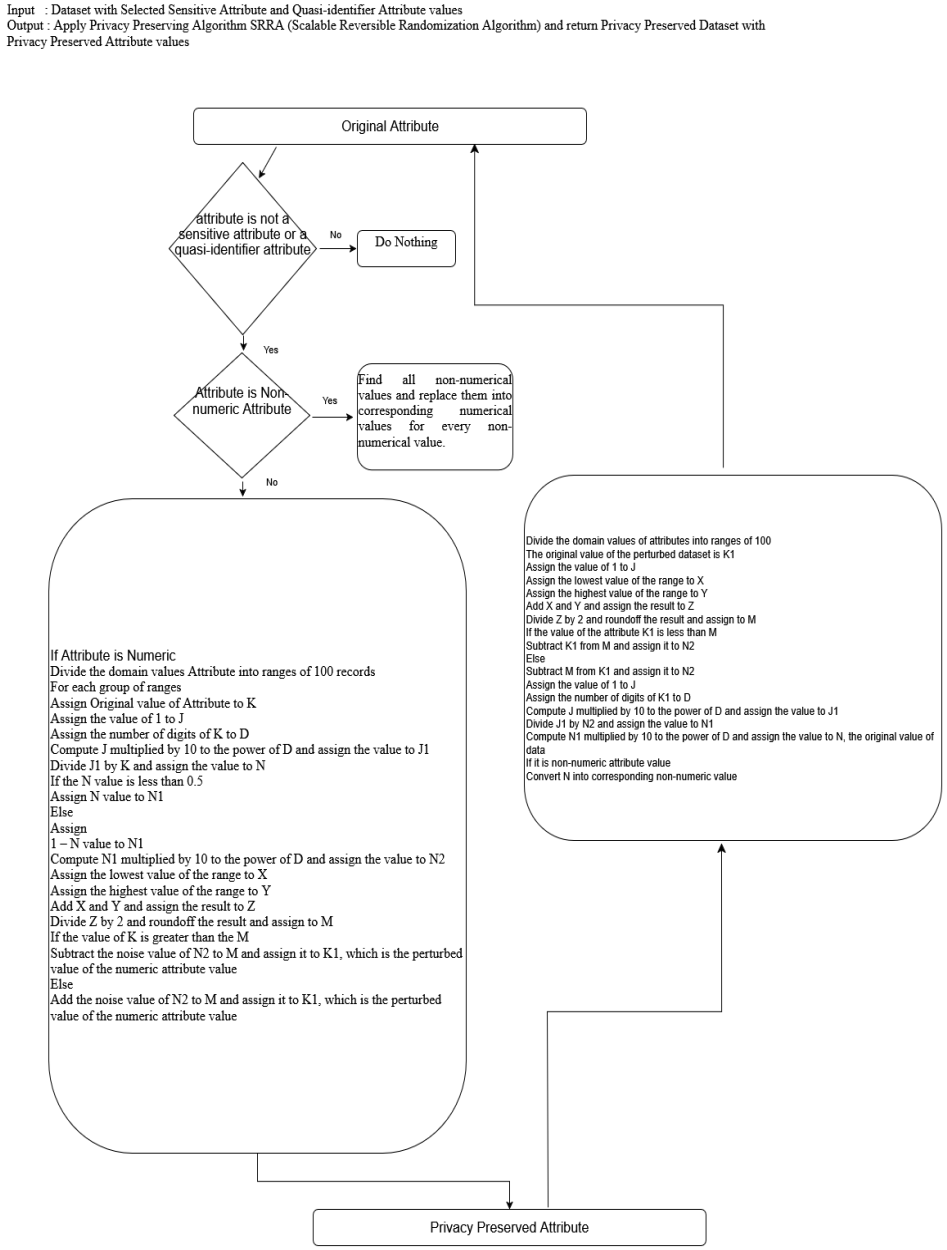

Fig. 1. Pseudocode of the proposed reversible privacy preserving algorithm of SRRA.

The selected sensitive and quasi-identifier attributes are perturbed by the proposed SRRA for privacy preservation. The application of SRRA in big data, which contains numeric and non-numeric data converted into definite sizes or formats, is suitable for the application of feature selection methods. The feature selection algorithm applied on the privacy conserved dataset, which is an important preprocessing step, will select an almost similar set of attributes, as the privacy conserving perturbation does not much affect the statistical characteristics of the experimental dataset.

The privacy-preserved dataset is used for big data analytics to get insights and hidden patterns from it. The application of SRRA results in the change in statistical characteristics of attribute values of the experimental dataset is negligible with less information loss, leading to better stability of feature selection and improved big data analytics results in terms of accuracy and error rate, along with robust privacy preservation of individuals' records. The SRRA processing time is reduced to around 15% in comparison to other similar methods in big data scenarios. The reversible privacy conserving algorithm of SRRA can be used to get the original dataset from the privacy-preserved dataset.

IV. Experiments

A. Methodology

The methodology of research is diagrammatically shown in the block diagram, in Fig. 2.

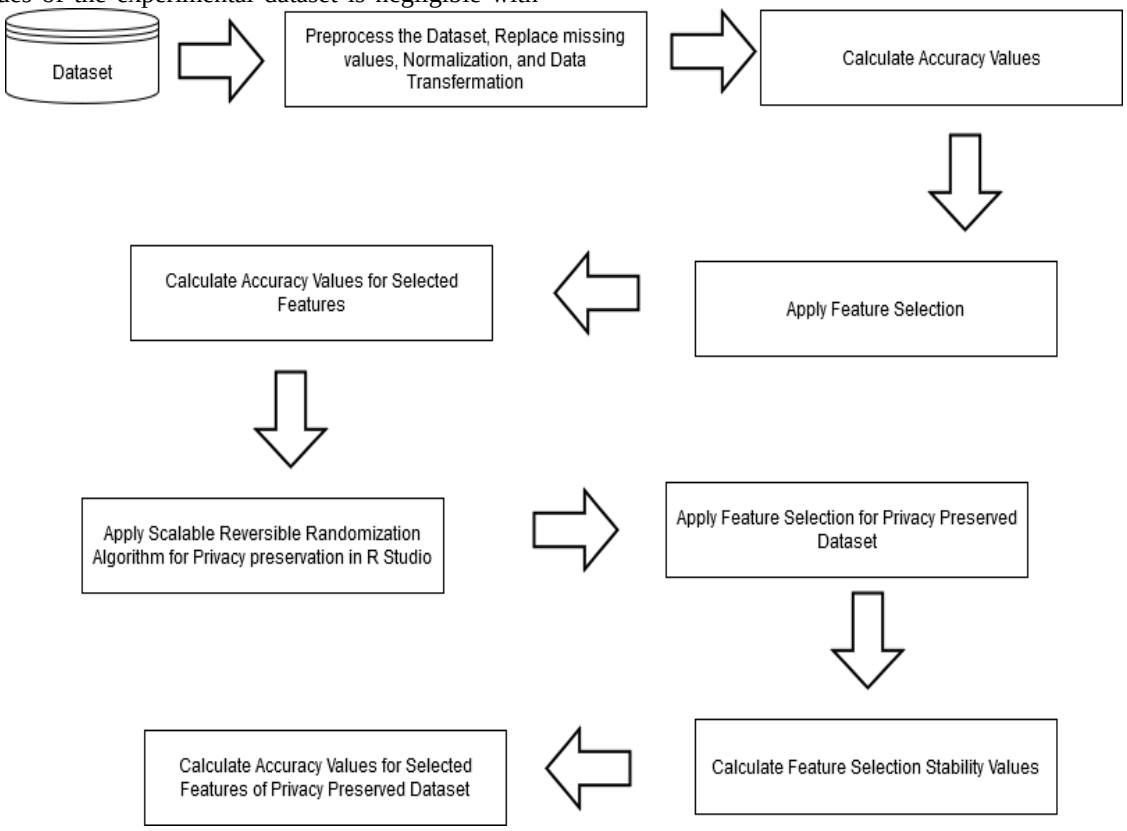

Fig. 2. Methodology.

B. Experimental Setup

The experiments are carried out in R-Studio. The application of privacy preserving modification by SRRA is implemented as a preprocessing step on the experimental datasets. The privacy-preserved datasets are harnessed in the Hadoop Distributed File System. The Hadoop distributed processing is implemented with three data nodes and one name node. The dataset used in the experiments is the Census Income dataset of the UCI repository.

C. Normalized Cardinality Penalty NCP

The metric to measure the information loss due to the manipulations of privacy-preservation is Normalized cardinality penalty (NCP) [22]. Information loss is an important parameter for the performance of the privacy preserving big data analytics algorithm. NCP is the metric of information loss in percentage, and hence it is suitable for big data scenarios. NCP for an equivalence class G for numeric attribute A is defined in Eq. (1):

$$NCP_A(G) = \frac{MAX_A^G - MIN_A^G}{MAX_A - MIN_A} \quad (1)$$

where, numerator MAX and MIN represent the range of attributes of class G, and denominator MAX and MIN represent the entire table.

D. Correlation-Based Feature Selection CFS

By seeing the degree of redundancy among them, along with the distinctive prophetic capability of each feature, the value of the attribute subsets is evaluated by CFS. The subsets of features that are prominently correlated within the class, with the inter-

correlation among classes, will be scanter preferences are taken by CFS [23]. Authors have GA, as a search method, with CFS as a fitness function. CFS selects the best feature subset and can be supplied with alternative search mechanisms. In Eq. (2), CFS is specified.

$$r_{zc} = \frac{k\, r_{zi}}{\sqrt{k + (k-1)\, r_{ii}}} \quad (2)$$

where, k refers to the number of subset features, r_{ii} refers to the mean inter-correlation between subset features, r_{zi} refers to the mean correlation between subset features and the class variable, and r_{zc} provides the correlation between other subsets of features and the class variable [23].

E. Dice's Coefficient

In [24], the Dice coefficient is the similarity measure used to measure the stability of feature selection, i.e., to calculate the overlap between two subsets. It included the Jaccard index. Dice takes a value between zero and one, where zero suggests no overlap and one suggests that the two subsets are identical. Dice, Jaccard, and Tanimoto are measures of similarity among them to calculate fluctuation in feature selection results. The Dice's coefficient is represented in the given Eq. (3):

$$\text{Dice}(F'1, F'2) = \frac{2\,|F'1 \cap F'2|}{|F'1| + |F'2|} \quad (3)$$

F. Results

In the experiments, four algorithms from Table I are used. Dice's coefficient is a metric used to measure fluctuations in feature selection results. The feature selection algorithm used in the experiment is CFS. Table I shows the relationship between changes in the physical characteristics of the dataset by privacy preserving perturbation effects on information loss, the fluctuations in feature selection results, and the accuracy of big data analytics outcomes.

In [20], privacy preservation is implemented in the HDFS distributed file system environment of real-time huge datasets. A secure map-reduce layer is introduced in MapReduce distributed processing for privacy preserving implementation. The methodologies used are generalization, slicing, anonymization (k-anonymity followed by l-diversity), and a one-way hash function, as privacy view generation in a parallel processing environment of distributed processing. In this method, the l value is kept constant for changing values of k for different anonymization views and processing time combinations for optimum configuration values. However, by using NCP, it is proven that information loss in 48% results in decreased data utility.

The algorithm is a combination of techniques like slicing, generalization, k-anonymity, l-diversity, one way hash function resulting in better privacy-preservation, but implementing this algorithm in the Hadoop distributed file system is very complex and time consuming, resulting in increased processing overhead in big data scenario. Stability of feature selection is 0.72 as it is drastically reduced along with data utility value for overall accuracy as 63.27% and accuracy of selected features as 68.59%.

TABLE I PERFORMANCE COMPARISON OF DIFFERENT TECHNIQUES

Privacy Preserving Technique Ref. No.	Information loss by NCP	Feature Selection Stability	Overall Accuracy		Accuracy of Selected Features	
			Before Privacy Preservation	*After Privacy Preservation*	*Before Privacy Preservation*	*After Privacy Preservation*
[20]	48%	0.72	74.51%	63.27%	79.68%	68.59%
[19]	35%	0.86	74.51%	69.68%	79.68%	75.51%
[18]	41%	0.77	74.51%	65.89%	79.68%	73.23%
[17]	59%	0.67	74.51%	61.87%	79.68%	65.86%
SRRA	26%	0.91	74.51%	72.28%	79.68%	77.41%

The privacy preserving algorithm by the authors in [19] is an improved scalable l-diversity approach, which has a higher stability of feature selection value of 0.86 and accuracy of overall feature as 69.68% and accuracy of selected features as 75.51% but it has lower privacy preservation compared to other methods. The information loss by NCP is 35%. The algorithm by [18] can be applied for semi-structured data using tagging techniques followed by scalable k-anonymity. The model is optimum in stability value of feature selection as 0.77, information loss by NCP as 41% and data utility of overall features as 65.89% and for selected features as 73.23%.

The algorithm by [17] is a cryptographic method, and it has the lowest stability of feature selection value of 0.67 and an overall accuracy value of 61.87% along with the accuracy of selected features as 65.86%, as it uses encryption methods, but it has better privacy preservation. In this method, the information loss by NCP is 59%, which is the highest among the listed methods.

The SRRA algorithm in [21] uses a randomization technique of calculating noise addition. The algorithm is applicable for both numeric and non-numeric data and hence suitable for big data. If the data is categorical or non-numeric, the attribute values are converted into equivalent numeric values before applying the algorithm. The proposed algorithm SRRA has the highest stability of feature selection value of 0.91 compared to other algorithms, an overall accuracy of 72.28%, and an

accuracy of selected features of 77.41%, which is higher than other algorithms. The information loss by NCP is also a minimum value of just 26%. This is because the change in statistical properties by SRRA is minimal compared to other algorithms, as the change in physical characteristics of the trial dataset is positively related to fluctuations in feature selection results and accuracy.

G. Discussion

The method in [20] is superior in terms of safeguarding private sensitive data. This method is very complex to implement in terms of implementation and computing cost, i.e., execution time. Implementation of the algorithm in [19], in the Hadoop distributed file system, using the MapReduce algorithm, is also easy compared to other models. This model in [18] is difficult to implement compared to other models used in the experiments.

Encryption methods in [17] are difficult to implement in big data scenarios, and data utility is also drastically reduced. So, there will be a relationship among privacy preserving perturbation, information loss, fluctuations in feature selection results, and accuracy.

As the characteristics of the dataset are related to the fluctuations in feature selection results and accuracy, the statistical characteristics of the dataset have minimal modification because of the privacy-conserving alteration by SRRA in [21]. Hence, SRRA give better results in terms of stability of feature selection, accuracy values of privacy-preserved dataset in terms of overall accuracy, and accuracy of selected features.

Information loss by NCP is also minimum, compared to other similar methods, as the change in characteristics of the dataset is minimal. Another advantage of SRRA is its suitability in big data scenarios, as the privacy preserving algorithm is a simple calculated noise addition technique and can be implemented in the preprocessing stage itself with less overhead, i.e., processing time, compared to similar privacy preserving algorithms without compromising on privacy-preservation.

V. Conclusion

The SRRA has higher stability of feature selection and accuracy values because the change in characteristics of the physical properties of the experimental dataset is minimal. However, the algorithm has better privacy preservation as there is a trade-off among variation in physical characteristics of the dataset, information loss, efficiency, fluctuations in feature selection results, as it is related with data utility, i.e. accurate data analytics results.

In the proposed privacy preserving big data analytics algorithm SRRA, both quasi-identifiers and sensitive features are anonymized by the randomization technique of calculating noise addition efficiently with minimal practical overheads in the preprocessing stage itself, and hence, because of the practical advantages, the implementation overhead is minimal compared to similar other algorithms in big data and cloud environments.

General privacy preserving algorithms are not working in a big data scenario. Hence, developing privacy-preservation algorithms in a big data scenario is very much challenging because of the implementation overhead and hence there will be a tradeoff among lesser information loss, better privacy preservation, improved stability of feature selection as it is directly linked with data utility, as selected features are used for data analytics for better efficiency, i.e. processing time and improved accuracy of data analytics results.

References

[1] Salem Alelyani, Huan Liu., The Effect of the Characteristics of the Dataset on the Selection Stability, IEEE DOI 10.1109/International Conference on Tools with Artificial Intelligence, 2011.167, 1082-3409/11, http:// ieeexplore.ieee.org/document/ 6103458, 2011.

[2] Salem Alelyani, Zheng Zhao, Huan Liu., A Dilemma in Assessing Stability of Feature Selection Algorithms, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications, 2011.99, 978-0-7695-4538-7/11, http:// ieeexplore.ieee.org/document/6063062, 2011.

[3] Salem Alelyani, On feature selection stability: a data perspective, Doctoral Dissertation, Arizona State University, AZ, USA, ISBN: 978-1-303-02654-6, ACM Digital Library, 2013.

[4] Barbara Pes, Feature Selection for High-Dimensional Data: The Issue of Stability, Proceedings of the 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2017), June 21–23, 2017.

[5] Jundong Li, Huan Liu, Challenges of Feature Selection for Big Data Analytics, Special Issue on Big Data, IEEE Intelligent Systems, eprint arXiv:1611.01875, 2017.

[6] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu, Feature Selection: A Data Perspective. ACM Comput. Surv, 50, 6, Article 94, 45 pages. DOI: https://doi.org/10.1145/3136625, 2018.

[7] C. Dwork, Differential privacy, Proceedings of the 33rd international conference on Automata, Languages, and Programming - Volume Part II, Pages 1–12 https://doi.org/10.1007/11787006_1pages 1-12, July 2006.

[8] Bayardo RJ, Agrawal A. Data privacy through optimal k-anonymization. In: Proceedings 21st international conference on data engineering, 2005 (ICDE 2005). Piscataway: IEEE; 2005.

[9] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. New York: ACM; 2005.

[10] Machanavajjhala A et al. L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd international conference on data engineering (ICDE'06), 2006. Piscataway: IEEE; 2006.

[11] Anil Prakash, Ravindar Mogili, Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 1, Issue 8, pp:28-33, 2012.

[12] Tiancheng Li, Jian Zhang, Ian Molloy.: Slicing: A New Approach for Privacy Preserving Data Publishing. IEEE Transaction on KDD, 2012.

[13] Aggarwal CC, Philip SY. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining. Springer: US; p. 11–52., 2008.

[14] Jiang R, Lu R, Choo KK. Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. Future Gen Comput Syst. 78:392–401., 2018.

[15] Mehta, B.B., Rao, U.P., Privacy preserving big data publishing: a scalable k anonymization approach using MapReduce. IET Software 11, 271–276. https://doi.org/10.1049/iet-sen.2016.0264., 2017.

[16] Mehta, B.B., Rao, U.P., 2018. Toward scalable anonymization for privacy preserving big data publishing. Recent Findings Intell. Comput. Tech. 708, 297–304.https://doi.org/10.1007/978-981-10-8636-6. Proceedings of the 5th ICACNI 2017, vol. 2., 2017.

[17] H. Shekhawat, S. Sharma and R. Koli, "Privacy-Preserving Techniques for Big Data Analysis in Cloud," Second International Conference on

Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, pp. 1-6, doi: 10.1109/ICACCP.2019.8882922, 2019.

[18] B. Mehta, U. P. Rao, R. Gupta and M. Conti, "Towards privacy preserving unstructured big data publishing", J. Intell. Fuzzy Syst., vol. 36, no. 4, pp. 3471-3482, 2019.

[19] Brijesh B. Mehta, Udai Pratap Rao, "Improved l-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing", Journal of King Saud University – Computer and Information Sciences 34, 1423–1430, 2022.

[20] Ganesh Dagadu Puri and D. Haritha, "Implementation of Big Data Privacy Preservation Technique for Electronic Health Records in Multivendor Environment" International Journal of Advanced Computer Science and Applications (IJACSA), 14(2), (DOI): 10.14569/IJACSA.2023.0140214, 2023.

[21] Mohana Chelvan P, Rajavarman V N, "The Scalable Reversible Randomization Algorithm (SRRA) for better privacy preservation, improved feature selection stability, and higher accuracy in Big Data Analytics" Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, E-ISSN: 1817-3195, Volume 103, Issue 03, pp. 794 – 801, 2025.

[22] Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N., Fast data anonymization with low information loss. In: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB '07, VLDB Endowment. VLDB Endowment, Vienna, Austria, pp. 758–769. https://doi.org/10.1109/ICDE.2007.369025., 2007.

[23] Mark A. Hall, Correlation-based feature selection of discrete and numeric class machine learning, Dept of Computer science, University of Waikato, 2000.

[24] Lei Yu, Chris Ding, and Ste1ven Loscalzo, Stable feature selection via dense feature groups. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 803–811, New York, NY, USA, 2008.