Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6402274, 12 pages
https://doi.org/10.1155/2022/6402274

Research Article

Big Data Privacy Preservation Using Principal Component Analysis and Random Projection in Healthcare

Ritu Ratra,[1] **Preeti Gulia**,[1] **Nasib Singh Gill**,[1] **and Jyotir Moy Chatterjee**[2]

[1]*Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India*
[2]*Department of Information Technology, Lord Buddha Education Foundation, Kathmandu, Nepal*

Correspondence should be addressed to Jyotir Moy Chatterjee; jyotir.moy@lbef.edu.np

Received 12 May 2022; Revised 23 June 2022; Accepted 28 June 2022; Published 9 August 2022

Academic Editor: Tahir Mehmood

With the rising usage of technology, a tremendous volume of data is being produced in the current scenario. This data contains a lot of personal data and may be given to third parties throughout the data mining process. Individual privacy is extremely difficult for the data owner to protect. Privacy-Preservation in Data Mining (PPDM) offers a solution to this problem. Encryption or anonymization have been recommended to preserve privacy in existing research. But encryption has high computing costs, and anonymization may drastically decrease the utility of data. This paper proposed a privacy-preserving strategy based on dimensionality reduction and feature selection. The proposed strategy is based on dimensionality reduction and feature selection that is difficult to reverse. The objective of this paper is to propose a perturbation-based privacy-preserving technique. Here, random projection and principal component analysis are utilized to alter the data. The main reason for this is that the dimension reduction combined with feature selection would cause the records to be perturbed more efficiently. The hybrid approach picks relevant features, decreases data dimensionality, and reduces training time, resulting in improved classification performance as measured by accuracy, kappa statistics, mean absolute error and other metrics. The proposed technique outperforms all other approaches in terms of classification accuracy increasing from 63.13 percent to 68.34 percent, proving its effectiveness in detecting cardiovascular illness. Even in its reduced form, the approach proposed here ensures that the dataset's classification accuracy is improved.

1. Introduction

Everyone is extremely data-centric in today's digital age. Data has been ubiquitous in recent years, and the size of data has increased dramatically. According to a recent poll published in April 2017 by the Storage Newsletter, the total data of the world was 16.1 Zettabytes (1021 bytes) in 2016, and is anticipated to increase to 163 Zettabytes by 2025. Data is generated by a variety of sources and organizations, including healthcare, banking,e-commerce, defense, insurance, social sites, and many more. Big Data is the term for this type of information. Big Data, refers to vast and massive datasets with a broad, diversified, and complicated structure that are challenging to store, analyze, and visualize [1]. Due to its size, velocity, and diverse nature, Big Data is defined as data that is massive in size and too complicated to be efficiently handled by traditional data management tools and approaches. Information extraction from Big Data is a major issue. Data from data warehouses and repositories can be mined for patterns that can be used for decision-making and analysis. This is referred to as Data Mining [2]. Knowledge Discovery in Databases is a method for extracting useful knowledge or information from data obtained from diverse sources (KDD). Big Data Mining, like Data Mining, allows users to apply analytics to large data sets in order to identify intriguing and useful patterns and knowledge. Big Data incorporates real data, such as personal information, from which any type of information about a specific person can be gleaned. To control the sharing and use of this data, policies and procedures must be established in order to prevent privacy violations. Privacy is one of the main issues with big data mining. Sensitive data from various applications is included in big data. Due to the risk of data exploitation and misuse, it is crucial to prevent unauthorized individuals from accessing and disclosing sensitive data. So, mining datasets distributed across many parties without leaking extra private information has become an important topic recently. Privacy preservation of information in big data is now a major challenge for the data mining process. Many people are now concerned that their personal data will be disclosed and exploited. They believe that their personal data should be kept private. Furthermore, there should be certain mechanisms in place to protect personal information [3, 4]. To address this problem, privacy-preserving data mining technologies have been demonstrated and implemented. PPDM technology, which permits data mining without revealing the characteristics of the original data, is becoming increasingly important. People from different phases of KDD have different privacy concerns. Only necessary and required data should be given to the data collector, according to the Data Provider who provides data from various sources. The data collector gathers information from information sources and stores it in databases and data warehouses. It is necessary to modify this information to make it secure without altering its relevance or significance. The data miner analyzes data from databases and data warehouses and mines it to uncover insightful patterns and rules. He must keep these private results hidden from unreliable parties. Decision-makers that use the outcomes of data mining for additional analysis and decision-making claim that it [5]. Various data mining approaches are combining security assurance systems that have been built based on a variety of irritations. PPDM contributes to the protection of sensitive information and personal data of individuals [6].This article seeks to implement perturbation methods in order to ensure information preservation. Values of dataset records have been altered in data perturbation methods to prevent the recovery of original dataset values. It also keeps the dataset's beneficial characteristics. Swapping, condensation, randomized response, additive noise, and other techniques preserve the dataset's features. There are some techniques that experiment to retain record-level patterns. It is done by replacing all sensitive data with some substitute data that are alike to those of records with alike non-sensitive data. It can be performed either by the distributions of sensitive data in the existence of particular non-sensitive data or can be done by using the mean of sensitive data from an accumulated group of records. Some different approaches, like geometric data perturbation and random projection, preserve the pair-wise distances of the dataset's records. Due to this, these methods are becoming more applicable in performing different data mining tasks in which predictions of particular records is performed like classification and regression [7–9]. These conversions and transformations are normally applied on numeric data. Furthermore, straightforward transformations are involved in these approaches. The proposed research introduces a privacy-preserving approach that is based on perturbation. This approach has been tested on healthcare datasets. The main contributions of this paper are as follows:

(I) The significance of privacy in data mining is discussed in this study.

(II) The current study presents a thorough analysis of various perturbation-based privacy preservation methods.

(III) This paper includes the analysis of two machine learning techniques: random projection for dimension reduction and principal component analysis for feature selection.

(IV) The main objective of this paper is to design and implement an Improved Random Projection Perturbation (IRPP) technique by combining two machine learning techniques-principal component analysis and random projection. This technique is used to perturb the dataset, before the data mining process.

(V) The proposed technique has been used to alter the cardiovascular and hypothyroid disease datasets in order to preserve privacy.

(VI) This paper provides an evaluation and comparison of the proposed technique with different machine learning classifiers - ANN, J48, and Naïve Bayes.

(VII) The proposed IRPP technique selects important features, reduces data dimensionality, and shortens training time, all of which lead to enhanced classification performance as assessed by accuracy, kappa statistics, mean absolute error, and other metrics.

Perturbation techniques have been developed as a solution to provide confidentiality on users' data by converting it into an unimaginable and unpredictable form [10–13]. In Perturbation method, modification of data is performed by adding a small noise or by changing the structure of the original dataset. Framework of privacy persevering using Perturbation is shown in Figure 1. Data perturbation simply means allowing whoever receives the data to receive slightly altered data. As shown in the Figure 1(a) [14] and Figure 1(b), it is only authorized person who can modify the data [15, 16]. After then publish the data to analyst for data mining process.

Perturbation can be used to protect privacy in the following ways [17]:

(i) Rotational Perturbation Algorithm: This method twists the estimation of the two credits in the matrix while maintaining the significance of the worth.

(ii) Projection Perturbation Algorithm: This technique achieves perturbation by transferring the data value from a big to a small dimensions space. The shifting of the data value is done at random.

(iii) Geometric Perturbation Algorithm: The geometric perturbation approach is a hybrid strategy that combines revolution, interpretation, and strengthening the given information admire in the matrix to deliver the nature of information conservation for the most part for clusters.

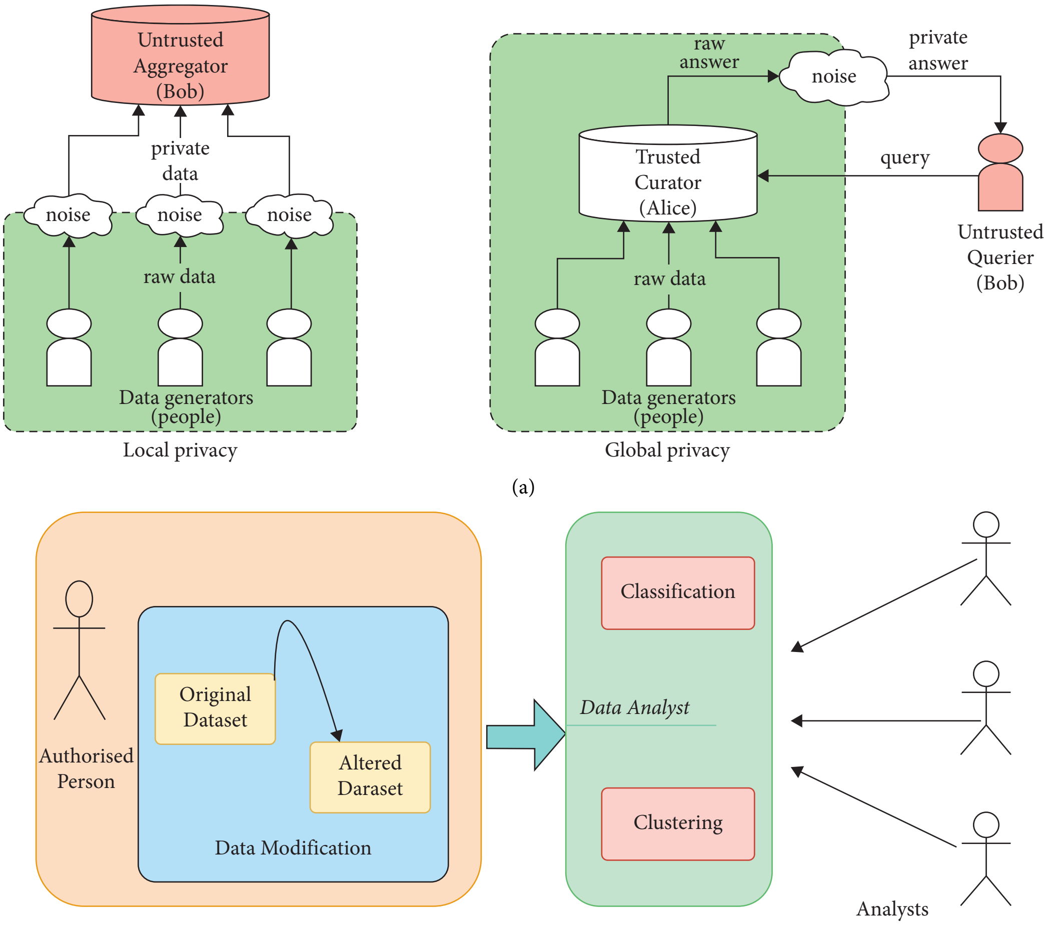


Figure 1: (a) Privacy Preservation using perturbation [14]. (b) Framework of privacy persevering using perturbation.

This research developed a strategy based on Random projection and principal component analysis that preserves the dataset's privacy by decreasing the high dimension to a low dimension while simultaneously improving data classification accuracy.

The rest of the paper is organized as follows: first, in Section 2, the research work on big data privacy preserving is explained. In Section 3, the various methods and approaches used in the proposed technique are discussed. The proposed approach is explained in Section 4. The experimental details of the suggested approach are presented in Section 5. In Section 6, the performance of the proposed technique is evaluated. Finally, Section 7 brought the proposed research work to a conclusion.

2. Literature Review

There has been a lot of research about the privacy and security of datasets. A variety of PPDM-related methodologies and procedures have been developed and applied in the past. However, most of these methods are not universally applicable. Mehta, et al. identified existing approaches from the literature of Natural Language Processing (NLP) to convert the unstructured data to structured form in order to apply k-anonymization over the generated structured records. A two-phase Conditional Random Field (CRF) based Named Entity Recognition (NER) approach is adopted to represent unstructured data into a structured form. They proposed an Improved Scalable k-Anonymization (ImSKA) to anonymize the well-represented unstructured data that achieves privacy-preserving unstructured big data publishing. The authors compared both of the proposed approaches namely NER and ImSKA with existing approaches and the results show that their proposed solutions outperformed the existing approaches in terms of F1 score and Normalized Cardinality Penalty (NCP), respectively [18]. Mariammala et al. (2021) presented a perturbation-based technique to preserve privacy in data mining. It is based on the additive rotation approach. They measured the level of privacy according to the variance of the original dataset. It was indicated that the protection of the original dataset has enhanced after applying their perturbation-based algorithm to that [19]. Chen and Omote proposed a privacy-preserving method based on perturbation. Their technique is based on dimensionality reduction. It is difficult to reverse while keeping the high usefulness of data. The authors combined dimensionality reduction algorithms with noise addition. Their approach is beneficial for privacy-preserving with high accuracy [20]. To improve the performance, Prasad Padavala, et al. combined deep learning models with big data technologies. Big data technologies are used to generate behavioral and content features. They used FCN, CNN, and RNN classifiers, as well as other deep learning models. The authors recognized attacks using the k-means clustering method in a multiclass model based on these selected features [21]. Sharma et al. proposed a hybrid method of privacy-preserving in data mining. Their proposed method is a combination of suppression and randomization. It is mentioned that this technique preserves data privacy and there is no information loss while regaining the original data value [22]. Rao et al. described "Synthesize Quasi Identifiers and apply Differential Privacy" (SQIDP). They proved that this approach is a more effective and scalable technique. They compared their algorithm (SQIDP) with the anonymization approach and stated that their algorithm is not prone to different types of attacks. It is also indicated through various experiments that SQIDP offered 100% data utility [23]. The feature selection fast correlation-based filter (FCBF) technique offered by Selvi and Pushpa was used to choose the important features and remove the duplicate data. To achieve data anonymization, they used k-anonymity on the dataset. To ensure that the final dataset is free of privacy exposure, an individual anonymization-based algorithm such as k-anonymity-related algorithm and differential privacy is presented [24]. According to Mary, the privacy level of the random projection approach is higher as compared to the other approaches. By using random projection the images can be highly preserved. With this technique, data can be more protective. The privacy level can be improved [25]. Binjubeir et al. provided an intensive review of existing PPDM approaches and classified various methods that are used for data modification. They explained the merits, and demerits, of the different PPDM techniques with the help of comparisons. This review study elaborated on the existing problems, challenges and some uncertain issues of PPDM [26]. According to Wang et al. , the methods which are based on Secure Multiparty Computation, are normally computationally expensive for use. Techniques that can be used on data streams must be exclusively designed for specific types of PPDM algorithms [27]. Mehta, et al. identified a few scalable approaches for Privacy-Preserving Big Data Publishing in literature, and the majority of them are based on k-anonymity and l-diversity. They proposed the Improved Scalable l-Diversity (ImSLD) approach which is the extension of Improved Scalable k-Anonymity (ImSKA) for scalable anonymization [28]. Pervez et al. described the different types of anonymisation techniques. These techniques are based on k-anonymity, t-closeness, etc. They applied these methods at the time of data mining and also performed anonymization at the time of merging records from different sources [29]. Prasanthi Kundeti and Chandra Sekhara Rao presented an effective hybrid technique for preserving the privacy of the dataset. In their approach, geometric data perturbation is used for numerical data and the k-anonymization technique (generalization) is used for categorical data [30].

Shobha Rani and Dhamodaran presented a survey on several techniques that impose privacy over the big data. The authors discussed the methodologies of Privacy-Preserving Cosine Similarity Computing protocol (PCSC) and Optimized Balanced Scheduling in detail (OBS). Cryptographic algorithms are at the foundation of their method. They employ a trapdoor function to make cost and time comparisons [31]. Geetha and Iyengar proposed a technique in which they performed perturbation with randomization. The data was generated in the form of intervals. They applied a classification algorithm to the modified data set and revealed that the accuracy of the dataset was well maintained [32]. Samir and Amin explored various Data perturbation methods. The data value of records gets perturbs in data perturbation methods. According to their study, the perturbation method is used to preserve the privacy of original values. Methods such as rank swapping, condensation, randomized response, and additive noise retained dataset properties [33]. Kumara Pandya, et al. concluded through their research that techniques such as geometric perturbation and random projection preserved the pair-wise distances between records. Using this makes them more helpful for data mining processes. Furthermore, he also described that these methods involved simple transformations of records. It was making them more effective to constantly use on data sets [34].

3. Methods and Algorithms

This section highlights the methods and algorithms that are used in the proposed technique. Two approaches are used here i.e. Random projection and principal component analysis.

3.1. Random Projection. Random projection has emerged as a powerful approach for dimensionality reduction. In random projection, the original high d-dimensional data are projected to a lower k-dimensional subspace by using a random $k \times d$ matrix. The key idea of random projection arises from the Johnson-Lindenstrauss (JL) lemma [35] The JL lemma shows that any set of s points in m dimensional space can embed into an O (log n/ε) - dimensional space such that the pairwise distances of any two points are maintained within an arbitrarily small factor. This property implies that it is possible to change the data's original form by reducing its dimension but still maintaining its statistical characteristics. Random projection is a way that is proposed to introduce the concept of random rotation. It is a method that prevents the required distances between data values. In this technique, the distance-preventing orthogonal matrix is replaced with the distance-preserving matrix. It also reduces the number of dimensions of the matrix to generate a projection on the dataset. Random projection can be of two types i.e. column-wise or row-wise. Column-wise projection is used to protect the records and it also preserves the distances between the records. Row-wise projection is used to preserve the features and the distance between the features. The random projection method is mainly applied to

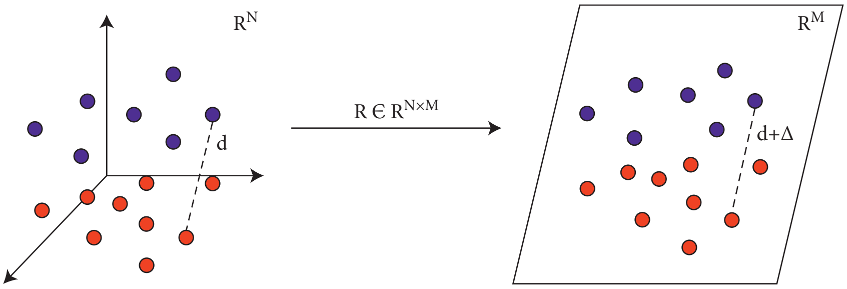


Figure 2: Dimension Reduction using Random Projection [37].

numeric data. Random projection methods are popular because they are powerful and simple. It is having low error rates as compared to other techniques. Random projection is worked on the JL lemma [36]. According to JL lemma "if points in vector space are projected onto a randomly selected subspace of suitably high dimensions, then the distances between the points are approximately preserved" as shown in Figure 2 [37].

The random projection has appeared as an effective method that is to reduce the dimension. In this technique, the original high-dimensional data are projected to a lower-dimensional space. It was suggested in 1984 which states that "A set of points in a high-dimensional space can be projected into a lower-dimensional subspace in such a way that relative distances between data points are nearly preserved". The lower dimension subspace that is used is selected randomly in the random projection technique.

3.2. Principal Component Analysis. The concept of feature selection has been broadly used in several areas of computer science. For example computer vision, machine learning. PCA (Principal Component Analysis) is also known as a feature selection technique and widely used method to reduce the dimensionality of datasets. It generally permits the complete technique to be used computationally more effectively. It is very helpful to enhance the accuracy of the technique and is usually supportive in improving the reduction of storage. PCA is usually helpful in improving the understanding of the dataset. It also helps in removing irrelevant features. Here PCA is selected as a feature selection tool to select the attributes according to the magnitude of their coefficients. Principal Component Analysis is a famous type of Unsupervised Learning Algorithm. As shown in the figure, the implementation of PCA is just like the Dimension Reduction Technique and the aim of PCA is to remove irrelevant and extraneous features of a dataset. PCA extracts the most dependent features contributing to the output. It has been broadly applied in a diversity of fields like recognition of patterns, compression of data files, data mining, and many more [38]. At the time of the Data mining process, raw data is provided for mining. Before interpreting the raw data, it has to make sure that some refinements are applied to the provided data. These refinements include preprocessing the data, for example removing the null values from the data. After that, the Feature selection technique is applied. This can be utilized in Principal Component Analysis (PCA) where the least contributing features are removed. In the last stage, the Data Transformations are applied. Here the user will apply different normalization methods to scale all the features in the same range.

The step by step procedure of PCA transformation [39] is given as follows:

Step1. Standardize the data (Center and scale).

To standardize the data given approach can be used.

$$Z = \frac{\text{value} - \text{mean}}{\text{Standard deviation}}. \tag{1}$$

According to this, subtract the mean and then dividing by the standard deviation for each variable.

When the standardization process is done, all the variables will be converted to the same scale.

Step2. Covariance Matrix Computation

This step is implemented to find to the relationship between various variables of dataset.

Step 3: Eigenvectors and Eigen values

In this step Eigenvectors and Eigen values of the Covariance Matrix are computed to recognize the principal components in order of significance.

Step 4: Feature Vector

In this step, more significant components (high Eigen values) are chosen and less significant (less eigenvalues) components are discarded. A matrix of vectors is formed with the remaining ones in known as a *Feature vector*.

Step 5: Recasting of the data along the Principal Components Axes

In the last step, feature vector is used to reconstruct the data. Reconstruction of data is performed from the original plane to the new one. This new plane is represented by the principal components. To achieve this, the transpose of the feature vector is multiplying with the transpose of the original data set, i.e.,

$$\text{FinalDataSet} = \left(\text{Feature Vector}^T * \text{Standardized Original DataSet}^T\right). \tag{2}$$

4. Proposed Technique

This section describes the proposed technique's design and analysis. The proposed privacy-preserving technique's framework is depicted in Figure 3. Data gathering and extraction, data preprocessing and preparation, feature selection, and dimension reduction are all part of the design phase. It is followed by the analysis phase which includes classification. The selection of the approach depends on the best outcomes obtained through the classification accuracy of provided perturbed dataset.

4.1. Designing of the Proposed Technique. The overall workflow of the proposed algorithm is shown in Figure 3. The entire framework of the paper is divided into two phases. These phases are:

Phase 1: This phase is concerned with the protection of the privacy of individuals in datasets. Basically, two modules are included in this phase. These are:

(a) *Feature Selection Module*: This is the module that is used for feature selection and to boost the accuracy of the classification technique. This is based on Principal component Analysis and applied to the original dataset before the Random Projection transformations of the dataset. The feature selection module is applied before two modules i.e. Random Projection module and the classification module. In this paper PCA-based feature selection approach is used for feature selection.

(b) *Random Projection Module*: Here, in this module, the perturbed data is modified again by using dimension reduction that is done with the help of the random Projection method. The random projection method is used to perturb the datasets. The accuracy of the perturbed dataset is also checked and compared with the original dataset.

Phase 2: This phase involves the classification process of perturbed datasets.

Classification module: After the implementation of the two modules mentioned above, perturbed data sets are being mined with some selected classification algorithms. Here, “ANN”, “NaiveBayed” and “J48” methods are taken as classification algorithms. Different matrices are also calculated on original datasets and compared with the perturbed dataset’s accuracy.

4.2. Algorithm for Proposed Technique. Step by step algorithm of the proposed technique is depicted in Figure 4 with the help of flow diaram.

Proposed technique: Improved Random Projection Perturbation (IRPP). (Algorithm 1)

The flow chart of the proposed technique is represented in Figure 4.

5. Implementation and Experiments

To evaluate the performance of the proposed technique, a number of comprehensive experiments are performed on different open-source datasets. This section introduces the experimental setups and discuss the evaluation results.

5.1. Experimental Setup. The proposed technique is implemented with the help of WEKA [40]. In these experimental analyzes, the proposed algorithm is implemented on the original datasets to get the altered datasets. Various measures such as accuracy, time is taken to build the model, kappa static measure, mean absolute error, and f-measure are determined on both datasets using various algorithms. These measures are used to examine the performance of the proposed technique on the projected dataset. To check the effectiveness and efficiency of the proposed technique, three classification algorithms are selected for implementation i.e. ANN, Naïve Bayes, and J48.

Datasets: In this paper, two datasets are used for experimental purposes [41, 42]. These datasets are the cardiovascular disease dataset and the hypothyroid disease dataset. The selection of datasets is based on their number of instances and number of attributes respectively. In the Cardiovascular disease dataset, instances are large and in the hypothyroid disease dataset, attributes are high. The main purpose of this paper is to prove that the proposed technique is applicable to both types of big datasets. The detail of each dataset in Table 1, which include the number of records, number of attributes, and description of attributes.

To implement the approach, first of all, PCA-based feature selection is applied to the original dataset. After that Random projection-based perturbation is applied to the modified dataset to get the final perturbed dataset. It is the property of PCA that all features are sorted from largest to smallest with regard to their Eigenvalues. Attribute selection is based on their rank. After that dimensions reduction is performed to get a perturbed dataset.

Figure 5 illustrates the cardiovascular and Hypothyroid datasets after perturbation respectively. It is observed that modified datasets are more secure as compared to the original dataset because it is difficult to access the altered data. So privacy can be preserved.

5.2. Evaluation Metrics. The performance of the proposed technique is evaluated using different classification metrics. These are accuracy, runtime, mean absolute error, Kappa statistics and F- measures [43].

(a) Accuracy: One parameter for assessing classification models is accuracy. The percentage of predictions our model correctly predicted is known as accuracy. The following is the formal definition of accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \tag{3}$$

Accuracy can also be determined in terms of positives and negatives for binary classification, as seen below:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \tag{4}$$

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

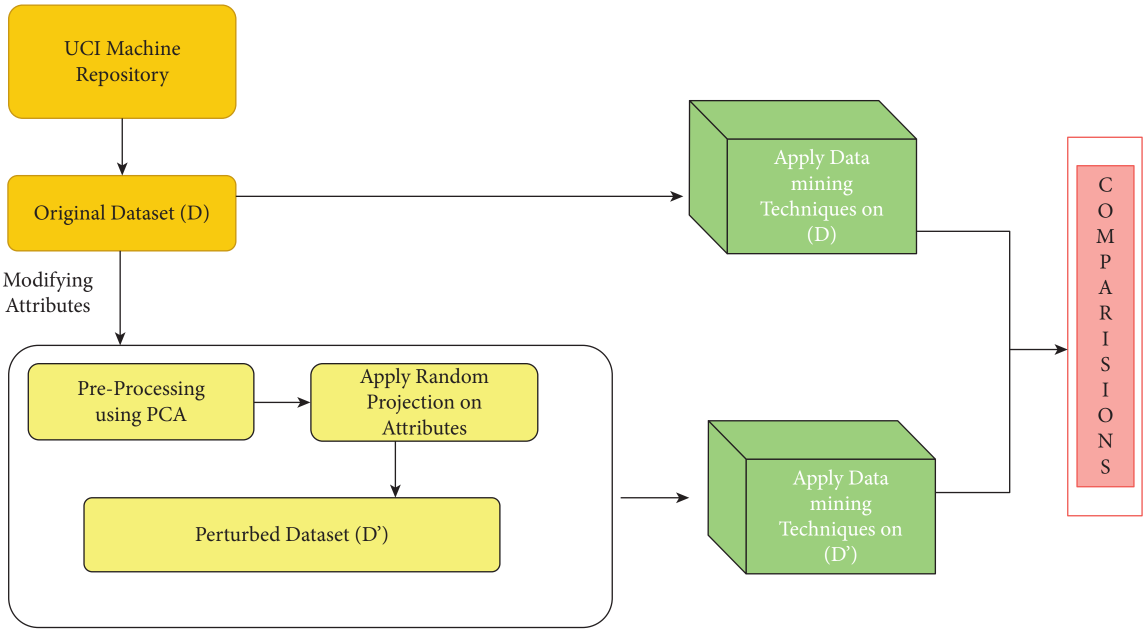


Figure 3: Overall Process flow of Proposed Algorithm.

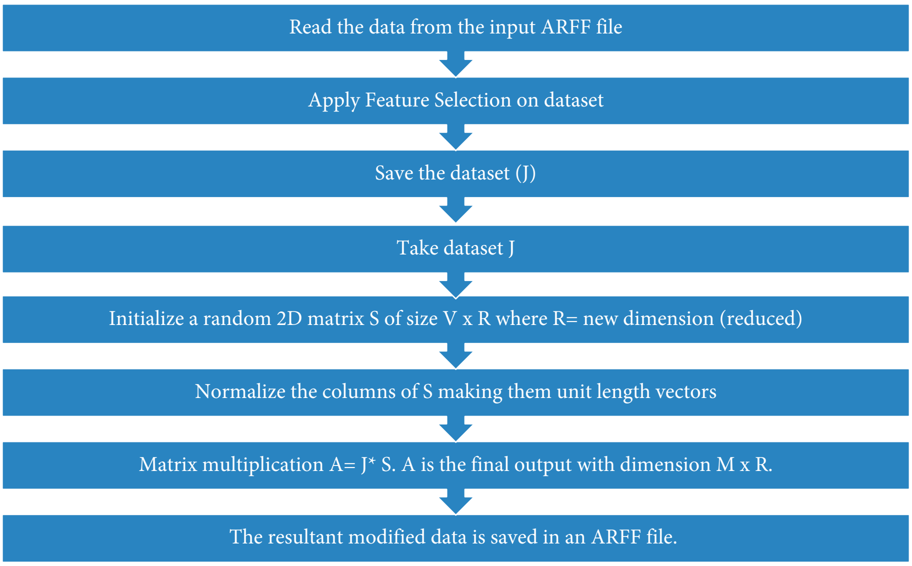

Read the data from the input ARFF file
Apply Feature Selection on dataset
Save the dataset (J)
Take dataset J
Initialize a random 2D matrix S of size V x R where R= new dimension (reduced)
Normalize the columns of S making them unit length vectors
Matrix multiplication A= J* S. A is the final output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

Figure 4: Flowchart of proposed algorithm.

(b) *Mean Absolute Error*: An assessment metric used with regression models is mean absolute error. The mean absolute value of each prediction error on each instance in the test set is the mean absolute error of a classifier with respect to the test set. The difference between the instance's true value and the expected value represents each prediction error.

(c) *Kappa Statistics*: A metric that differentiates observed accuracy and expected accuracy is the Kappa statistic (or value) (random chance). In addition to evaluating a single classifier, the kappa statistic is also used to assess classifiers amongst themselves. Additionally, it accounts for random chance (agreement with a random classifier), making it less deceptive than just using accuracy as a statistic.

$$\mathit{Kappa\ statistics} = \frac{(\text{observed accuracy} - \text{expected accuracy})}{(1 - \text{expected accuracy})}. \tag{5}$$

(d) *F-measures*: A combined measure for precision and recall calculated as

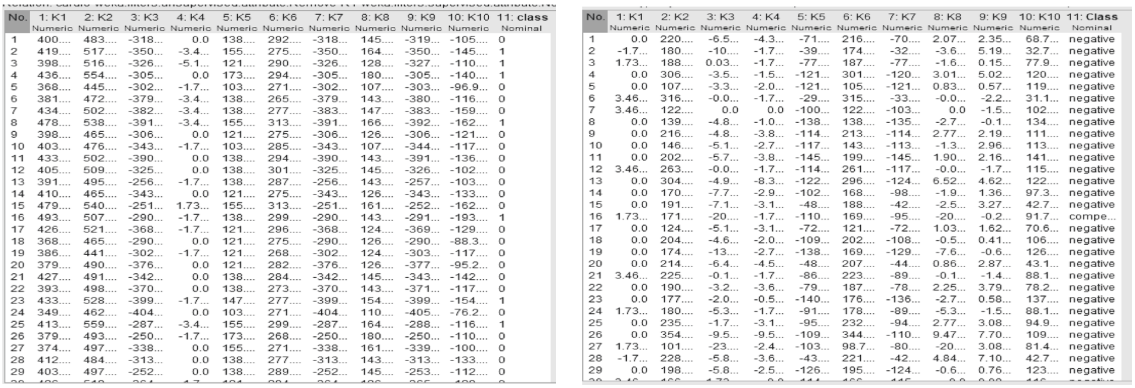
$$\frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}. \tag{6}$$

```
**Input**: ARFF File with original data values.
**Output**: ARFF File with perturbed data after applying Enhanced Random Projection Perturbation.
(1) Read the data from the input ARFF file of the dimension $M$ x N ($M$ = different samples and $N$= Original features) and named it P.
(2) Apply Feature Selection on dataset P by using Principal component Analysis and reduce the small ranked features.
(3) Save the dataset say $J$.
(4) Take dataset $J$, of the dimension $M \times$ V ($M$ = different samples and $V$ = Reduced features)
(5) Initialize a random 2D matrix S of size V × R where $R$ = new dimension(reduced)
(6) Normalize the columns of S making them unit length vectors.
(7) Matrix multiplication $A = J *$ S. A is the final output with dimension $M \times R$.
(8) The resultant modified data is saved in an ARFF file.
```

Algorithm 1:

Table 1: Description of datasets.

Name of dataset	Number of instances	Number 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



No.	1: K1 Numeric	2: K2 Numeric	3: K3 Numeric	4: K4 Numeric	5: K5 Numeric	6: K6 Numeric	7: K7 Numeric	8: K8 Numeric	9: K9 Numeric	10: K10 Numeric	11: class Nominal
1	400....	483....	-318...	0.0	138....	292....	-318...	145....	-319...	-105....	0
2	419....	517....	-350...	-3.4...	155....	275....	-350...	164....	-350...	-145....	1
3	398....	516....	-326...	-5.1...	121....	290....	-326...	128....	-327...	-110....	1
4	436....	554....	-305...	0.0	173....	294....	-305...	180....	-305...	-140....	1
5	368....	445....	-302...	-1.7...	103....	271....	-302...	107....	-303...	-96.9...	0
6	381....	472....	-379...	-3.4...	138....	265....	-379...	143....	-380...	-116....	0
7	434....	502....	-382...	-3.4...	138....	277....	-383...	147....	-383...	-159....	0
8	478....	538....	-391...	-3.4...	155....	313....	-391...	166....	-392...	-162....	1
9	398....	465....	-306...	0.0	121....	275....	-306...	126....	-306...	-121....	0
10	403....	476....	-343...	-1.7...	103....	285....	-343...	107....	-344...	-117....	0
11	433....	502....	-390...	0.0	138....	294....	-390...	143....	-391...	-136....	0
12	405....	509....	-325...	0.0	138....	301....	-325...	145....	-326...	-102....	0
13	391....	495....	-256...	-1.7...	138....	287....	-256...	143....	-257...	-103....	0
14	410....	465....	-343...	0.0	121....	275....	-343...	126....	-343...	-133....	0
15	479....	540....	-251...	1.73...	155....	313....	-251...	161....	-252...	-162....	0
16	493....	507....	-290...	-1.7...	138....	299....	-290...	143....	-291...	-193....	1
17	426....	521....	-368...	-1.7...	121....	296....	-368...	124....	-369...	-129....	0
18	368....	465....	-290...	0.0	121....	275....	-290...	126....	-290...	-88.3...	0
19	386....	441....	-302...	-1.7...	121....	268....	-302...	124....	-303...	-117....	0
20	379....	490....	-376...	0.0	121....	282....	-376...	126....	-377...	-95.2...	0
21	427....	491....	-342...	0.0	138....	284....	-342...	145....	-343...	-142....	0
22	393....	498....	-370...	0.0	138....	273....	-370...	143....	-371...	-117....	0
23	433....	528....	-399...	-1.7...	147....	277....	-399...	154....	-399...	-154....	1
24	349....	462....	-404...	0.0	103....	271....	-404...	110....	-405...	-76.2...	0
25	413....	559....	-287...	-3.4...	155....	299....	-287...	164....	-288...	-116....	1
26	379....	493....	-250...	-1.7...	173....	268....	-250...	180....	-250...	-110....	0
27	374....	497....	-338...	0.0	155....	271....	-338...	161....	-339...	-100....	0
28	412....	484....	-313...	0.0	138....	277....	-313...	143....	-313...	-133....	0
29	403....	497....	-252...	0.0	138....	289....	-252...	145....	-253...	-112....	0

No.	1: K1 Numeric	2: K2 Numeric	3: K3 Numeric	4: K4 Numeric	5: K5 Numeric	6: K6 Numeric	7: K7 Numeric	8: K8 Numeric	9: K9 Numeric	10: K10 Numeric	11: Class Nominal
1	0.0	220....	-6.5...	-4.3...	-71....	216....	-70....	2.07...	2.35...	68.7...	negative
2	-1.7...	180....	-10....	-1.7...	-39....	174....	-32....	-3.6...	5.19...	32.7...	negative
3	1.73...	188....	0.03...	-1.7...	-77....	187....	-77....	-1.6...	0.15...	77.9...	negative
4	0.0	306....	-3.5...	-1.5...	-121...	301....	-120...	3.01...	5.02...	120....	negative
5	0.0	107....	-3.3...	-2.0...	-121...	105....	-121...	0.83...	0.57...	119....	negative
6	3.46...	316....	-0.0...	-1.7...	-29....	315....	-33....	-0.0...	-2.2...	31.1...	negative
7	3.46...	122....	0.0	0.0	-100...	122....	-103...	0.0	-1.5...	102....	negative
8	0.0	139....	-4.8...	-1.0...	-138...	138....	-135...	-2.7...	-0.1...	134....	negative
9	0.0	216....	-4.8...	-3.8...	-114...	213....	-114...	2.77...	2.19...	111....	negative
10	0.0	146....	-5.1...	-2.7...	-117...	143....	-113...	-1.3...	2.96...	113....	negative
11	0.0	202....	-5.7...	-3.8...	-145...	199....	-145...	1.90...	2.16...	141....	negative
12	3.46...	263....	-0.0...	-1.7...	-114...	261....	-117...	-0.0...	-1.7...	115....	negative
13	0.0	304....	-4.9...	-8.3...	-122...	296....	-124...	6.52...	4.62...	122....	negative
14	0.0	170....	-7.7...	-2.9...	-102...	168....	-98....	-1.9...	1.36...	97.3...	negative
15	0.0	191....	-7.1...	-3.1...	-48....	188....	-42....	-2.5...	3.27...	42.7...	negative
16	1.73...	171....	-20....	-1.7...	-110...	169....	-95....	-20....	-0.2...	91.7...	compe...
17	0.0	124....	-5.1...	-3.1...	-72....	121....	-72....	1.03...	1.62...	70.6...	negative
18	0.0	204....	-4.6...	-2.0...	-109...	202....	-108...	-0.5...	0.41...	106....	negative
19	0.0	174....	-13....	-2.7...	-138...	169....	-129...	-7.6...	-0.6...	126....	negative
20	0.0	214....	-6.4...	-4.5...	-48....	207....	-44....	0.86...	2.87...	43.1...	negative
21	3.46...	225....	-0.1...	-1.7...	-86....	223....	-89....	-0.1...	-1.4...	88.1...	negative
22	0.0	190....	-3.2...	-3.6...	-79....	187....	-78....	2.25...	3.79...	78.2...	negative
23	0.0	177....	-2.0...	-0.5...	-140...	176....	-136...	-2.7...	0.58...	137....	negative
24	1.73...	180....	-5.3...	-1.7...	-91....	178....	-89....	-5.3...	-1.5...	88.1...	negative
25	0.0	235....	-1.7...	-3.1...	-95....	232....	-94....	2.77...	3.08...	94.9...	negative
26	0.0	354....	-9.5...	-9.5...	-109...	344....	-110...	9.47...	7.70...	109....	negative
27	1.73...	101....	-23....	-2.4...	-103...	98.7...	-80....	-20....	3.08...	81.4...	negative
28	-1.7...	228....	-5.8...	-3.6...	-43....	221....	-42....	4.84...	7.10...	42.7...	negative
29	0.0	198....	-5.8...	-2.5...	-126...	195....	-124...	-0.6...	0.76...	123....	negative

Figure 5: Snapshots of Cardiovascular and Hypothyroid dataset after Enhanced Random Projection perturbation.

6. Comparison and Results

In this paper, a hybrid privacy-preserving algorithm is proposed and implemented. This technique experiments on various data sets and corresponding results were noticed. The experimental outcomes show.

That privacy-preserving using the proposed technique is the best as compared to traditional techniques because it has a higher value in accuracy, kappa statistic, and runtime. The proposed technique enhances the classification accuracy while preserving the privacy of the dataset. Conventional techniques are not suitable for large volume data sets. Different experiments illustrated that the proposed Random projection method with feature selection achieves better efficiency in big data than the simple classification method. Here privacy is preserved without any loss of accuracy. Comparison between existing and proposed techniques is shown in Tables 2 and 3.

Table 2 presents the runtime observations made by several classification techniques. It shows the rates of classification accuracy, runtime, kappa statistic, Mean Absolute Error, and F-measure on the original samples and perturbed samples obtained using the ANN, J48, and Naïve Bayes classifiers. It is observed that the feature selection and dimension reduction by the proposed technique can reduce the runtime of the classification algorithm. The proposed

Table 2: Performance Measure of Classification algorithm on the cardiovascular dataset.

Accuracy measurement	ANN		Naïve-bayes		J48	
	Classification approach (original dataset)	Classification approach (perturbed dataset)	Classification approach (original dataset)	Classification approach (perturbed dataset)	Classification approach (original dataset)	Classification approach (perturbed dataset)
Time taken to build (SEC)	113.43	88.3	0.25	0.24	15.39	10.8
Correctly classified instances (%)	63.71	68.83	58.85	59.68	72.99	73.01
Incorrectly classified instances (%)	36.28	31.16	41.14	40.31	27.01	26.89
Kappa statistic	0.2744	0.3767	0.1768	0.1936	0.4581	0.4599
Mean absolute error	0.4357	0.3943	0.4221	0.4193	0.3616	0.3701
F-measure	0.631	0.6880	0.6860	0.6890	0.7290	0.7360

Table 3: Performance Measure of Classification algorithm on the hypothyroid dataset.

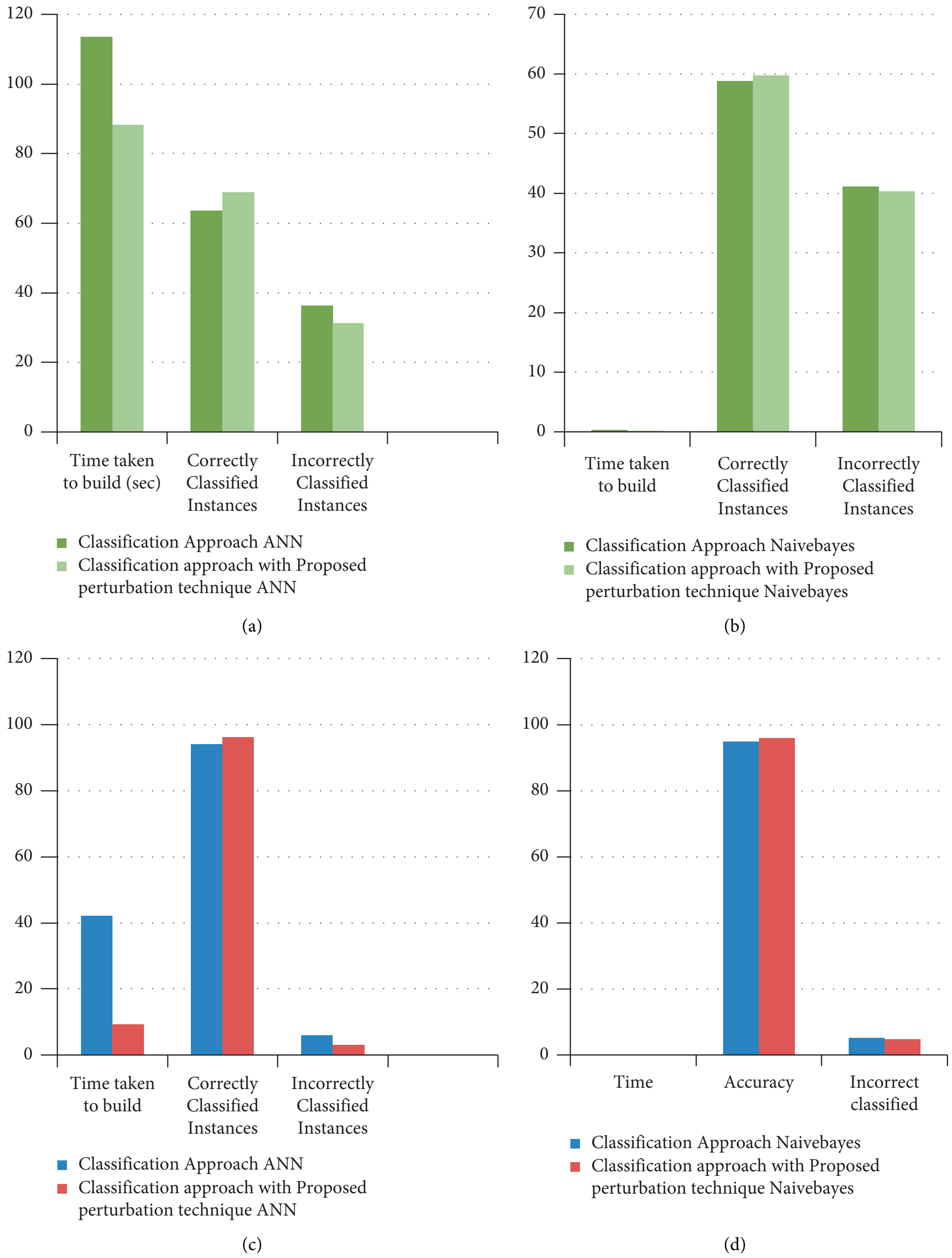
Accuracy measurement	ANN		Naive bayes		J48	
	Classification approach (original dataset)	Classification approach (perturbed dataset)	Classification approach (original dataset)	Classification approach (perturbed dataset)	Classification approach (original dataset)	Classification approach (perturbed dataset)
Run time	42.18 sec	9.28 sec	0.07 sec	0.01 sec	0.13 sec	0.04 sec
Correctly classified instances (%)	94.16	96.16	94.80	95.89	99.57	98.99
Incorrectly classified instances (%)	5.83	3.075	5.19	4,74	0.43	1.01
Kappa statistic (%)	0.5519	0.7630	0.543	0.624	0.9707	0.9532
Mean absolute error (%)	0.1522	0.017	0.033	0.026	0.003	0.004
F-measure (%)	0.971	0.986	0.975	0.989	0.998	0.998

technique of privacy-preserving increases the accuracy and kappa statistic rate as compared to the original samples of datasets.

For example, the classification accuracy rate on the original dataset's samples of the Cardiovascular dataset is 63.71% and the proposed algorithm achieves the accuracy rate of 68.83%. The runtime of the original dataset is 114.18 seconds and of the perturbed dataset is 88.3 seconds. As depicted in the table, it is observed that the proposed technique has better or almost same results in all measure than the traditional model of classification.

Table 3 shows the performance of the proposed privacy-preserving method to the conventional classification models on hypothyroid datasets. Here, the runtime, accuracy, kappa statistic measure, mean absolute error, and Root mean squared errors area are compared on the given datasets. It is portrayed in Table 3 and observed that the proposed technique has better results in all measures than the traditional model of classification.

Figure 6 depicts the graphical performance of the proposed privacy-preserving approach compared to existing classification techniques on different datasets. On cardiovascular datasets, it shows how the random projection-based privacy-preserving strategy outperforms the traditional classification methodology. As can be seen in the figures, the suggested method outperforms traditional classification algorithms in terms of accuracy, runtime, kappa statistics, Mean absolute error, and F- measures. As depicted in Figure 6(a), it is observed that the proposed approach has better accuracy measures than the traditional model of classification using ANN classifier on the Cardiovascular dataset. In the instance of the Naive-Bayes classification method, Figure 6(b) shows that the proposed technique outperformed. It has a greater accuracy of 59.68 percent than traditional ones. It has an accuracy of 59.68 percent which is higher than the traditional one. On the Hypothyroid dataset, Figure 6(c) shows a performance analysis of the suggested technique employing an ANN classifier. In comparison to the original dataset, it also takes less time to run. As seen in the graph, the suggested IRPP technique takes 9.28 seconds, which is significantly less than the 42.18 seconds required by the original dataset. The performance study of the suggested technique on the Hypothyroid dataset utilizing the Naïve-Bayes classifier is



(a)

(b)

(c)

(d)

Figure 6: Performance analysis of enhanced random projection perturbation based privacy-preserving model to the conventional model on cardiovascular dataset and hypothyroid dataset using ANN and Naive Bayes.

depicted in Figure 6(d). When compared to the original dataset, it has a high level of accuracy, less runtime, and high F- Measures.

7. Conclusion

This paper proposed and implemented an effective privacy-preserving algorithm based on perturbation for healthcare datasets. Different classification techniques have also been used in various investigations. Feature selection and dimension reduction are the foundations of the proposed technique. Precise perturbation has been achieved here by combining Random Projection with the Principal Component Analysis approach. On different large datasets, the usefulness and accuracy of the proposed technique have been tested in classification algorithms- ANN, Naive Bayes, and J48 classifiers. With the use of diverse experimental results, it has been determined that the suggested privacy-preserving technique is more accurate and efficient than traditional techniques. In the case of cardiovascular datasets, the suggested technique outperforms traditional techniques in terms of runtime, accuracy, efficiency, mean absolute error, kappa statistic measure, and F-measurer, even after the perturbation. In the case of the hypothyroid dataset, experimental outcomes on all measurements (efficiency, accuracy, runtime, kappa statistic measure, and mean absolute error value) are better or almost identical to the previous approach model.

Data Availability

Data will be available on request from the submitting author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Sagiroglu and D. Sinanc, "Big Data: A Review," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, San Diego, CA, USA, May 2013.

[2] S. Shelke and B. Bhagat, "Techniques for privacy preservation in data mining," *International Journal of Engineering Research*, vol. 4, no. 10, 2015.

[3] R. Ratra and P. Gulia, "Privacy preserving data mining: techniques and algorithms," *International Journal of Engineering Trends and Technology*, vol. 68, no. 11, pp. 56–62, 2020.

[4] M. El-hasnony, H. M. El Bakry, and A. A. Saleh, "Comparative Study among Data Reduction Techniques over Classification Accuracy," *International Journal of Computer Applications*, vol. 122, no. 2, pp. 8–15, 2015.

[5] Morioh, "Principal Component Analysis," https://morioh.com/p/847c472d1146.

[6] S. Kumar Bhandare, "Data distortion based privacy preserving method for data mining system," *International Journal of Emerging Trends & Technology in Computer Science*, vol. 2, no. 3, 2013.

[7] M. Naga Lakshmi and K. Sandhya Rani, "SVD Based Data Transf Ormation Methods for Privacy Preserving Clustering," *International Journal of Computer Applications*, vol. 78, no. 3, 2013.

[8] S. Nagendra Kumar, R. Aparna, "Sensitive attributes based privacy preserving in data mining using k-anonymity," *International Journal of Computer Application*, vol. 84, no. 13, pp. 1–6, 2013.

[9] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "Comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, no. 1, pp. 1–36, 2016.

[10] M. Reza Keyvanpour and S. Seifi Moradi, "Classification and evaluation the privacy preserving data mining techniques by using a data modification–based framework," *International Journal on Computer Science and Engineering, ISSN*, vol. 3, no. 2, pp. 862–871, 2011.

[11] Li Liu, M. Kantarcioglu, and B. Thuraisingham, "The Applicability of the Perturbation Model-Based Privacy Preserving Data Mining for Real-World Data," in *Proceedings of the 6th IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, Hong Kong, China, December 2006.

[12] N. Rajesh, K. Sujatha, and A. A. Lawrence, "Survey on privacy preserving data mining techniques using recent algorithms," *International Journal of Computer Application*, vol. 133, no. 7, pp. 30–33, 2016.

[13] P. Gulia and C. Hemlata, "Privacy preserving data mining of vertically partitioned data in distributed environment- an experimental analysis," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 10, pp. 2973–2987, 2018.

[14] 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, pp. 1–18, 2018.

[15] T. Revathi and N. Ramaraj, "Data privacy preservation using data perturbation techniques," *International Journal of Soft Computing and Artificial Intelligence*, ISSN, vol. 5, no. 2, pp. 10–13, 2017.

[16] R. Ratra, P. Gulia, and N. S. Gill, "Evaluation of Re-identification risk using anonymization and differential privacy in healthcare," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, 2022.

[17] N. Sangavi, "Data pertubation techniques in privacy preserving data mining," *International Journal of Research and Advanced Development (IJRAD)*, ISSN, vol. 4, no. 2, pp. 2581–4451, 2020.

[18] B. Mehta and U. P Rao, "Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1423–1430, 2022.

[19] S. Mariammala, S. Dr, A. Kavithamanib, and S. Baradhwajc, "An additive rotational perturbation technique for privacy preserving data mining," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 9, pp. 2675–2681, 2021.

[20] Z. Chen and K. Omote, "A Privacy Preserving Scheme with Dimensionality Reduction for Distributed Machine Learning," in *Proceedings of the 2021 16th Asia Joint Conference on Information Security (AsiaJCIS)*, pp. 45–50, Seoul, Republic of Korea, August 2021.

[21] S. Prasad Padavala, S. Tamilarasu, and S. Gurava, "Big data feature selection model for intrusion detection using data analytics," *International Journal Of Engineering Research & Technology (IJERT) ICRADL*, vol. 9, no. 5, 2021.

[22] V. Sharma, D. Srivastava, and D. Soni, "Pramod kumar, A novel hybrid approach of suppression and randomization for privacy preserving data mining," *xIlkogretim Online - Elementary Education Online*, vol. 20, no. 5, pp. 2451–2457, 2021.

[23] P. R. M. Rao, S. M. Krishna, A. P. S. Krishna, and A. P. S. Kumar, "Novel algorithm for efficient privacy preservation in data analytics," *Indian Journal of Science and Technology*, vol. 14, no. 6, pp. 519–526, 2021.

[24] U. Selvi and S. Pushpa, *Big data feature selection to achieve anonymization*, pp. 59�67, Springer, Singapore, 2020.

[25] A. Mary, "A random projection approach to secure medical images," *International Journal of Advanced Research*, vol. 7, no. 3, pp. 1298–1301, 2019.

[26] M. Ahmed, A. A. Sadiq, M. A. B. Khurram Khan, A. S. Sadiq, and M. Khurram Khan, "Comprehensive survey on big data privacy protection," *IEEE Access*, vol. 8, Article ID 20067, 2020.

[27] J. Wang, C. Liu, X. Fu, X. Luo, and X. Li, "A three-phase approach to differentially private crucial patterns mining over data streams," *Computers & Security*, vol. 82, pp. 30–48, 2019.

[28] B. Mehta, U. P. Rao, R. Conti, and M. Conti, "Towards privacy preserving unstructured big data publishing," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 4, pp. 3471–3482, 2019.

[29] F. Pervez and M. D. Ahmed, "Prediction of collapse potential for gypseous sandy soil using ANN technique," *Journal of Engineering Science & Technology*, vol. 15, no. 2, pp. 1236–1253, 2020.

[30] N. Prasanthi Kundeti and M. V. P. Chandra Sekhara Rao, "A combined approach for privacy preserving classification mining," *International Journal of Pure and Applied Mathematics*, vol. 14, no. 1, pp. 188–194, 2019.

[31] P. Shobha Rani and V. Dhamodaran, "Security and privacy in big data analytics," *International Journal on Intelligent Electronic Systems*, vol. 10, no. 2, pp. 32–35, 2016.

[32] M. A. Geetha and N. Iyengar, "Non-additive random data perturbation for real world data," *Procedia Technology*, vol. 4, pp. 350–354, 2012.

[33] S. Patel and K. R. Amin, "Privacy Preserving Based on PCA Transformation Using Data Perturbation Technique," *International Journal of Computer Science & Engineering Technology*, vol. 4, no. 5, 2013.

[34] B. KumarPandya, U. kumar Singh, and K. Dixit, "A robust privacy preservation by combination of additive and multiplicative data perturbation for privacy preserving data mining," *International Journal of Computer Application*, vol. 120, no. 1, pp. 28–31, 2015.

[35] M. Gupta, "Random projection for dimension reduction moving beyond PCA," 2022, https://medium.com/data-science-in-your-pocket/random-projection-for-dimension-reduction-27d2ec7d40cd.

[36] E. Bingham and H. Mannila, "Random Projection in Dimensionality Reduction: Applications to Image and Text Data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, San Francisco, CA, USA, August 2001.

[37] K. Kun Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92–106, 2006.

[38] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in *Proceedings of the 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, Yichang, China, November 2010.

[39] Built In, "Data-Science," https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

[40] R. Ratra and P. Gulia, "Experimental evaluation of open source data mining tools (WEKA and orange)," *International Journal of Engineering Trends and Technology*, vol. 68, no. 8, pp. 30–35, 2020.

[41] D. Dua and C. Graff, "UCI machine learning repository, Thyroid Disease Data Set," https://archive.ics.uci.edu/ml/datasets/thyroid+disease.

[42] S. Ulianova, "Cardiovascular Disease dataset, version 1," 2022, https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

[43] A. Mishra, "Metrics to evaluate your machine learning algorithm," 2022, https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.