Research Article

Rupali Tajanpure* and Akkalakshmi Muddana

Data analysis with performance and privacy enhanced classification

https://doi.org/10.1515/jisys-2022-0215
received June 28, 2022; accepted March 02, 2023

Abstract: Privacy is the main concern in cyberspace because, every single click of a user on Internet is recognized and analyzed for different purposes like credit card purchase records, healthcare records, business, personalized shopping store experience to the user, deciding marketing strategy, and the list goes on. Here, the user's personal information is considered a risk process. Though data mining applications focus on statistically useful patterns and not on the personal data of individuals, there is a threat of unrestricted access to individual records. Also, it is necessary to maintain the secrecy of data while retaining the accuracy of data classification and quality as well. For real-time applications, the data analytics carried out should be time efficient. Here, the proposed Convolution-based Privacy Preserving Algorithm (C-PPA) transforms the input into lower dimensions while preserving privacy which leads to better mining accuracy. The proposed algorithm is evaluated over different privacy-preserving metrics like accuracy, precision, recall, and $F1$-measure. Simulations carried out show that the average increment in the accuracy of C-PPA is 14.15 for Convolutional Neural Network (CNN) classifier when compared with results without C-PPA. Overlap-add C-PPA is proposed for parallel processing which is based on overlap-add convolution. It shows an average accuracy increment of 12.49 for CNN. The analytics show that the algorithm benefits regarding privacy preservation, data utility, and performance. Since the algorithm works on lowering the dimensions of data, the communication cost over the Internet is also reduced.

Keywords: convolution, privacy preservation, feature reduction, data utility, communication cost

1 Introduction

This section briefly explains the overview and privacy preservation of data mining process, motivation of the research, contribution of the research, and organization of the research paper.

1.1 Overview

Data mining and analysis have influenced everyone's life. You start searching any term on google and end by selecting various auto-search word combinations and getting suggestions based on your search query. You go for online shopping and many recommendations like "users who purchased this, also viewed these items." With this, one can experience personalized shopping. Many websites ask to register and log in before giving access to the data needed by the user. Here, everyone wants to keep their information secure.

*** Corresponding author: Rupali Tajanpure,** Computer Science and Engineering, GITAM School of Technology, GITAM University, Hyderabad, 502329, Telangana, India, e-mail: tajanpure.rupali@kbtcoe.org
Akkalakshmi Muddana: Computer Science and Engineering, GITAM School of Technology, GITAM University, Hyderabad, 502329, Telangana, India, e-mail: amuddana@gitam.edu

Due to this, fake data are shared by users on the Internet affecting data analytics. Extracting valuable knowledge from data while preserving the privacy of data has become essential now [1]. In Industry 4.0, data privacy is very important while sharing data for analytics [2]. When the data collection is huge and needs to be mined, it requires high-capacity servers and storage, which gives rise to the issue of real-time mining. Vast data storage as well as sharing of information over the Internet brings out different issues of data security. It also suffers from communication costs. So, some fast and privacy-preserving data processing techniques are needed for analysis and classification as hackers are also there to hack the sensitive information of users despite using security protocols and encryption. In such situations, privacy is defined as keeping the personal as well as social information of individuals in a secure manner when there is a presence of a third party who performs the computation based on their details [3]. Consequently, the next subsection explains the procedure of privacy-preserving process and motivation of this research work.

1.2 Privacy-preserving data mining (PPDM) and motivation

PPDM methods have been established to permit the abstraction of information from datasets while avoiding the revelation of delicate data or data subject's identities. Two or more researchers can collaborate on a single dataset with the PPDM technique. In a multilateral environment, sensitive and confidential information about dataset can be obtained by performing a data mining process without disclosing the data of each party to other parties. In PPDM, it is necessary to protect sensitive information by either masking data or changing the data while maintaining classification accuracy. The PPDM algorithms are judged based on the rate of unique data recovery from modified data, loss of information, and an effect on accuracy [4]. Domadiya and Rao proposed privacy-preserving association rule mining for handling the issue related to the healthcare system in IoT applications [5]. In this COVID-19 pandemic, the healthcare data of COVID patients are collected for ministry decisions, to know the requirement of oxygen beds and the requirement of medicine supply in coming days, etc. Here, patients are reluctant to share their data. The real challenge in data collection is to project on the privacy of patients. It is again necessary to provide privacy to the data while protecting the classification accuracy of sensitive data, which is a very cumbersome task [6]. Usually, data are transformed to preserve its privacy. It may reduce the usefulness of data. The effect of the privacy preservation algorithm is calculated by means of the risk of data disclosure from deployed data. The degree of privacy preservation is expressed as the variance of recovered and original deployed data. The characteristics of PPDM are like preserving the privacy of sensible information and the accuracy of data mining, and it should not compromise the access to sensitive data [7].

1.3 Research contribution

In this research, a data security method is developed and it eases data classification by reducing the computational time and storage space requirements while retaining the accuracy of classification. Since original data get transformed into another form, they become secure. This also proves very useful as no one can hack or make misuse of individual's data. The data remain unrecognizable and still give accuracy in mining. This article contributes a privacy-preserving method, which works very effectively on the level of privacy preservation, data hiding, and classification accuracy preservation while reducing the complexity of data processing. The highlights of this research are as follows:

- The proposed technique uses an overlap technique that adds the transformation to get output in a reduced format. Since input gets processed in an overlapped fashion, technique uses FFT algorithms which need less time.
- Data utility and mining accuracy are maintained despite privacy preservation, and data are transformed into reduced form leading to less storage and computation time.

- Data transformation is irreversible so one can share transformed data without security concerns.
- With the help of the proposed C-PPA method, the classification algorithm gives more accurate results in the classification process.

Additionally, the novelty of this research is explained here. Generally, in machine learning algorithms, adding many more features at first will give a more accurate result. However, the output of model will reduce after a certain level, with the rising number of elements. This is due to the high dimensionality of features presented in the dataset. This problem arises due to the exponential decrease of samples with increasing dimensionalities. The dimensionality of the feature space expands more and more when the features must be added without increasing the number of training samples. And, this leads to the problem of overfitting in machine learning algorithms. This is reduced with the help of the proposed C-PPA technique which reduces the features and obtains the collection of principle features. In this research, the proposed feature reduction algorithm gives novelty, and by this algorithm, the classifier achieves better classification accuracy.

1.4 Organization of the study

This study is categorized as follows. The next section elaborates on existing research works in the domain under consideration. The third section focuses on the proposed method of privacy preservation and privacy-preserving against member inference attacks. The fourth section describes the simulation outcomes of proposed privacy-preserved data, and the last section concludes the paper.

2 Literature survey

Privacy preservation is a vast area to work on. In literature, randomization of data is suggested, which adds noise to mask some sensitive attribute values. Here, experts add sufficient noise to hide the data, but this may drive a loss of accuracy. In k-anonymity, generalization and suppression techniques are used to hide the individuality of the records while l-diversity is an improvement over k-anonymity. A privacy-preserving data mining algorithm for perturbing the original data which is applicable to all data types for an arbitrary probability distribution is proposed by Ge et al. [8].

While using an online social network, trust evaluation plays an important role. Fatehi et al. proposed an AI and graph-based hybrid model for the same which gives an improvement in accuracy reaching 95% of present trusted paths [9]. Privacy-preserving approach is used in different application areas like IoT, machine learning [10,11], deep learning [12], and distributed machine learning.

The k-anonymity method is affected by the dimensionality of the data. For high dimensionality, it becomes tough to hide the individuality of the data without loss of information. Also, the curse of high dimensionality affects privacy preservation in the data mining process [13]. When a data holder wants to share private information regarding health care or banking data for research purposes, one should guarantee that the personal data shared will remain unidentified. In literature, k-anonymity-based systems like Datafly, μ-Argus, and k-Similar are proposed [14]. The problem of optimal k-anonymity is NP-hard [15]. Maintaining data utility after alteration in data is also an important thing to focus on [16]. The results of data mining applications are changed by quashing some rules of associative rule mining, and it is considered one of the techniques to preserve privacy. In query auditing, the results of a query are modified. When data are distributed among multiple nodes, some cryptographic techniques or protocols are applied to maintain data privacy.

Process mining uses event data to improve processes. But here, event data contain sensitive information. Majid Rafiei et al. [17] proposed a group-based privacy preservation approach that focuses on some

interpretable and adjustable parameters to take care of different privacy aspects. An algorithm to integrate two or more open government data into one data set to make the mining process effective is proposed by Jae Lee et al. [18]. This method also allows the user to set the threshold level of privacy, which will help the user to balance well between data utility and data disclosure risk.

Mariana Cunha et al. [19] introduced privacy-preserving mechanisms for the privacy preservation of users. The author presents the existing study on heterogeneous data types with systematic analysis. To search for optimal feature set partitioning, Nissim Matatov et al. [20] proposed a data mining privacy by decomposition (DMPD) algorithm based on a genetic algorithm (GA). The k-anonymity-based method is used to evaluate the classification performance of the proposed methodology on ten datasets. A review of privacy preservation in machine learning with differential privacy is presented by Gong et al. He presented an approach balancing privacy and the utility of data [21]. A big data classification and security approach is proposed by Hababeh et al. [22,23] to achieve high data mobility in the cloud system. A detailed survey of different network security attacks, their categorization, and typical network attacks in that category is done by Jing et al. [24] with their performance in terms of detection scalability and flexibility.

This differential privacy-preserving approach is also used in Distributed machine learning [25]. A review of the application, challenges, opportunities, and metrics of industrial data privacy for differential privacy in Industrial IoT is thoroughly put in ref. [26] by Jiang et al. For IoT-based applications, a privacy-preserving data scheme is introduced by Almagrabi and Bashir [27], and it is based on the trust score of available resources. In healthcare systems, due to centralized storage and data control, the conventional cloud and client server-based modules suffer from single-point failure. Sharma et al. [28] proposed a distributed data management system-based blockchain approach which is useful for solving such problems. Privacy leakage problem is faced for collinear data, and Zhang et al. [29] proposed a correlation reduction technique based on feature selection. The results show improvement in data utility.

The large number of features in clinical records increases the data communication cost over the network [30]. So, it is desirable if the privacy preservation algorithm also works for feature reduction. Skubalska-Rafajłowicz [31] worked on a new method based on Gaussian random projection, and it is very easy as well as protects the privacy of image data [31]. Nazir et al. [32] used an auto-encoder for anomaly detection for the SCADA network. Auto-encoder generates a reduced representation of data. An input layer in a neural network is followed by a hidden layer with a smaller number of nodes and hence reduced representation of data, and this process is known as encoding. Likewise, the hidden layer to output layer data transmission is known as decoding [32].

To enable data exchange in IoT, Vehicular ad hoc networks (VANETs) link two or more vehicles wirelessly. Location privacy of VANET is considered a top most priority as it contains a crucial piece of information. Most of the existing papers do not address the threats in their research, so, in this research, Ahmed et al. [33] provide the summary of location privacy attacks and their resolutions to overcome the issues in IoT environment. Also, digital signature method-based cryptographic solutions are explained in this research. VANET transforms public transport into a safer wireless network to enhance safety and efficiency. The communication in a network is performed with the help of different kinds of nodes like vehicles, Roadside Units (RSUs), traffic signals, and other wireless communication devices.

Security threats are increasing day by day, so it is very essential to develop security algorithms to control the threats. In this research, Junejo et al. [34] proposed different artificial intelligence (AI) methods for RSUs. A comparison between trust and cryptography was presented in this research which was based on the applications and requirements of VANET. In general, IoT is susceptible to different identities like attacks and threats and these are controlled thanks to the growth in consumer's density with low power access nodes. For transfer conditions, Memon et al. [35] discovered the possible flaws associated with IoT security situations, and a new technique is developed to perceive a spoofing attack that enquires the probability distributions of received power for mobile users. An algorithm named MTFLA is developed to guarantee detection and protection in a huge sensitive region that means the developed algorithm is used in the maximized chance of an attack.

A range of extremely delicate personal features like age, gender, sexual orientation, intelligence, and personality traits of people are automatically and easily accessed by the unknown person through

Facebook likes. Analysis of this research [36] was conducted based on a dataset which contains 58,000 volunteers' details like detailed demographic files, Facebook likes, and results of various psychometric tests. The pre-processing is done using a dimensionality reduction technique which pre-processes the Likes data and uses a logistic/linear regression algorithm to predict individual psychodemographic profiles. Online personalization and privacy implication are done with the help of association between the likes and attributes. A machine learning classifier was utilized by the attacker to deduce a target user's private information like political view, sexual orientation, and location details from public data like page likes and rate scores. The reason for this kind of inference attack was due to users who use vulnerable mobile and web applications. The existing method suffered from large utility loss of user's public data, so, in this research, Jia and Gong [37] developed a method named as AttriGuard which was a practical defense method against attribute inference attacks. By finding the probability distribution formulation, a constrained convex optimization problem was solved in this research.

Almaiah et al. [38] proposed a centralized system based on Industrial IoT to work on security and privacy in IIoT networks using blockchain-based deep learning framework. Duan et al. [39] use deep learning inference for cloud-based framework for security and privacy maintenance. Alkhelaiwi et al. [40] used privacy-preserving deep learning (PPDL)-based techniques to take care of privacy. He proposed a partially homomorphic encryption scheme, which processes confidential information without privacy leakage.

3 Proposed privacy preservation system design

The user surfing the Internet does many activities. For every user, their clickstream data, registration on different sites, and profile data are recorded. For this behavior, the data are collected and stored on the web server for different kinds of applications. The architecture of the proposed system is shown in Figure 1.

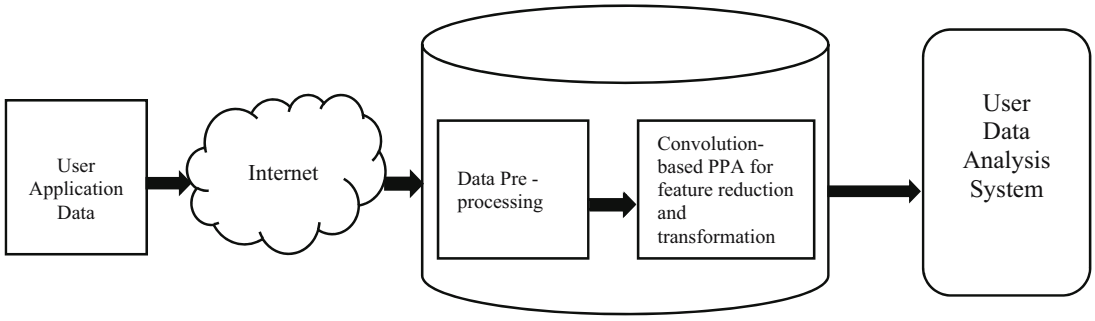

Figure 1: Architecture of the proposed system.

In this research, healthcare application is considered, and it utilizes the UCI repository dataset for performing the privacy preservation process. In healthcare applications, different kinds of parameters like patient's name, age, gender, and disease details are recorded and stored on the webserver for keeping the confidentiality of the records. Generally, the real-time data collection process takes more time to collect the data so, in this research, the publicly available UCI repository dataset is taken and it is processed with the help of webserver. Here, the data collected is large leading to more processing time. The webserver performs the pre-processing process on the collected data, and the pre-processing includes replacing missing data with column average and normalization technique. Finally, the proposed C-PPA algorithm is applied to perform the feature reduction technique on pre-processed data. After performing the feature reduction process, the user data analysis system performs the classification. The process flow of the proposed C-PPA methodology is described in the next subsection.

3.1 Data pre-processing

Generally, pre-processing process is considered as a primary process that directly affects the success rate of proposed methodologies. Performing other processes without pre-processing step reduces the data complexity, since the real-world data are unclean. Likewise, the repeated data or duplicate data also decrease the training process of the proposed methodology. In order to overcome these issues, this research performs two different kinds of a process named data imputation and Min–Max normalization, and they are performed before applying the proposed feature reduction technique.

Let us consider the original value to be N and the normalized value tends to N'. The range of N can be given as [min, max], and the new range is represented as [min′, max′]. For the mapping of the original value from one range to a new range, the normalized value is given by,

$$N' = \frac{N - \min}{\max - \min}(\max' - \min') + \min'. \tag{1}$$

Using equation (1), Min–Max normalization has been calculated for all values in the dataset. Before going to feature extraction process, we have generated one example for the Min–Max normalization process. After converting words into numerical values, Min–Max normalization is applied. The strength of the Min–Max normalization procedure is that the correlation among the values of original data is maintained. The data redundancy can be diminished, and the uprightness is improved. After the pre-processing, the feature reduction technique is performed on the pre-processed data to diminish the dimensionality of the features. Moreover, the feature reduction method is detailed in the following subsection.

3.2 Proposed C-PPA technique

After performing the pre-processing steps, C-PPA is applied to the pre-processed data to obtain reduced features for classification purposes. Convolution is one of the operations in digital signal processing (DSP) applications which converts every single value in the input sequence by delta function or scaling and shifting of unit impulse. From this, it is observed that the dominant dimensions in the input result in output with the same dominance. The first step of C-PPA is to divide data tuples into groups, with each group containing elements nearly the same as the closest, 2^n. These groups act as input to C-PPA. After the application of C-PPA, the output contains a reduced set of elements. Basically, C-PPA is based on the circular convolution of the signals expressed as,

$$Y(L) = \sum_{n=0}^{N-1} x_1(n)x_2((L-n))_N \text{ for } L = 0, 1, 2, .., N-1. \tag{2}$$

The significant property of Discrete Fourier Transform (DFT) is the circular convolution which is expressed in equation (2). The multiplication of DFTs of two sequences shows the same results as the circular convolution of the same in the time domain [41–46]. The proposed C-PPA procedure is given below.

Algorithm 1: Convolution based privacy preserving method.
Input: input features for a tuple under consideration.
Output: $Y(n)$ - output features.
Begin
Step 1: Divide input features in two parts $x1(n1)$ and $x2(n2)$ containing nearly same elements ... $n1 \approx n2$
Step 2: Decide N as number of features in output such as N= nearest 2^r which is 2,4,8,16,32,64.....
Step 3: Add zeros to $x1(n1)$ and $x2(n2)$ so as to make number of elements in each sequence similar to N

$$x1(N) = [x1(n1),\ \text{zero pad}(1,\ N - n1)]$$
$$x2(N) = [x2(n2),\ \text{zero pad}(1,\ N - n2)]$$

Step 4: Convolve $x1(N)$ and $x2(N)$ using equation (1) and following steps
Step 5: $Y(K)$ = FFT(First part)⋆FFT(second part).....use point by point multiplication
Step 6: C – PPA output = IFFT($Y(K)$)
End

From the above algorithm, one can reach Y(K) by taking the FFT of the output of C-PPA. But the hacker will get stuck at the point-by-point multiplication stage as numerous possible combinations can be there. So, Step 5 in the algorithm is irreversible. One cannot recover the original data from the extracted data. Hence, convolution-based privacy-preserving algorithm is the loss method. Likewise, the observed reduction in features is more effective when applied to data tuples containing a large number of features. The calculation of DFT is performed using Fast Fourier Transform (FFT) algorithms which proves beneficial in terms of calculations and hence computation time due to the following reasons:

a. Computation with FFT algorithm – for N Number of values, DFT calculation needs N^2 complex multiplications whereas FFT calculation needs $(N|2 \log 2N)$. Similarly, DFT calculation needs $(N^2 - N)$ complex additions whereas FFT calculation needs $(N \log 2N)$. Hence, the speed improvement is observed. Hence, FFT is a fast algorithm to calculate the DFT and with FFT, and the convolution operation is carried with high speed. Convolution performed using FFT is also called as high-speed convolution.
b. Overlapped Computations – here, computations are performed in overlapped fashion and then, the results are added to get the final data sequence
c. The data tuple features present in input undergo convolution and are transformed into a new set of features containing a smaller number of features. The new features are nothing but a combination of input features. The dominance of features is preserved in output; hence, we get better classification accuracy of output.

With the proposed system architecture, the data will be collected only at the central server and processed there. Now, the main server can share the transformed data 'n' number of times with other servers for data analytics purposes, and it can be shared through any network. The data will remain safe and trustworthy for analysis. Also, the central web server can store only the transformed representation of data, i.e., output of C-PPA, as it is the reduced form and gives better accuracy than the original data. The proposed algorithm gives output in reduced dimensions, so it proves better in terms of computational time and space requirements. Likewise, it also has some advantages when compared to the perturbation-based methods which are given below:

- reduction in data;
- extraction of data in a different form so that original data can't be predicted;
- data privacy is achieved;
- data utility is maintained;
- data mining accuracy is preserved;
- reduction in computational time;
- reduction in storage space;
- improved accuracy of mining;
- reduced communication costs over the Internet.

In this proposed C-PPA algorithm, the data tuple can be broken into a group of features for parallelism. Every group is processed parallelly in a overlapped manner. The output of these small groups is computed fast and then output features are fitted one over another to get the final output. This is known as overlap-add C-PPA algorithm which transforms the input in some anonymized and reduced form. From the algorithm, it is clear that N is nearly similar to a number of features in each group. If we consider an example of Parkinson's disease, the total features in one tuple are 754. We divide features into two groups with 377 in each group. So, we decide N = 512 (nearest 2^r). Hence, the features in output after the application of C-PPA with N = 512 will be 512. Here, the output has reduced features. Also, the output is a transformed version of

the input which is not reversible. Now, the output of C-PPA can be shared with different servers for analysis without data leakage risk. After performing the feature reduction process, the classification is done by using different machine learning algorithms like like SVM, DT, CNN and NB classifiers. These classifiers achieve better classification accuracy by using the proposed feature reduction technique. Subsequently, the next subsection explains the privacy preserving against the member inference attacks.

4 Simulation setup and evaluation

The overall implementation and results are computed with the help of MATLAB 2015b tool. Likewise, Intel core i5-7200U CPU @ 2.50 GHz with 8 GB RAM and 64-bit Windows 10 Operating System laptop is used to perform the simulation process. The performance of the proposed feature reduction technique is analyzed by means of different performance parameters like accuracy, precision, recall, and *F*-measure. Likewise, the performance of the proposed C-PPA is compared with different classifiers like CNN, SVM, DT, and Naive Bayes (NB) with and without feature reduction techniques. The classification algorithms are implemented by ourselves to demonstrate the efficiency of the proposed C-PPA technique. Subsequently, the next subsection explained the dataset description

4.1 Dataset description

Datasets are taken from the UCI repository for the evaluation of the proposed algorithm. UCI is a large dataset repository which contains different kinds of datasets to perform various operations. From this repository, six disease datasets are selected, which are pre-processed by missing value imputation and Min–Max normalization technique. The proposed algorithm is evaluated on selected datasets. Table 1 shows the datasets under consideration with their dimensions before and after the application of C-PPA. The observations show that the C-PPA algorithm also plays a major role in feature reduction too. The reduction in features is possible due to the selection of N and division of total attribute of a tuple in different groups which is an inbuilt methodology of the C-PPA algorithm.

Table 1: Dataset details

Name of dataset	No. of tuples	No. of features	No. of features after applying C-PPA
Parkinson's Disease	756	754	512
Heart Disease	303	13	8
Kidney	402	25	16
Hepatitis	132	19	16
Br. Cancer	699	9	8
Arrhythmia	452	280	128

The binary classification datasets listed in Table 1 undergo the C-PPA algorithm. CNN, DT, SVM, and NB classifiers are selected for evaluation of the proposed algorithm as it does not require any domain knowledge for its construction. Also, its execution is fast and easy to understand for humans [33]. To check the usefulness of data for further analysis, the mining accuracy is recorded for each dataset after the application of all the classifiers.

Table 1 clearly displays that the original features of Parkinson's disease dataset which are reduced from 754 to 512. The total reduction for Parkinson's disease dataset is 242 features. Here, the C-PPA transforms

the original data dimensions into lower dimensions. The data is transformed or extracted in reduced form with the application of C-PPA. In this method, the important or high weighted features maintain their weight in output after convolution, so the accuracy of convolved data does not get reduced [42].

Bertino et al. [43] presented the desired dimensions for the evaluation of the PPDM algorithm. It mentions efficiency, scalability, level of privacy achieved, quality of data after application of PPDM, and data hiding failure [44–46]. All these parameters are essential during the assessment of the quality of the algorithm under consideration. The efficiency is determined in terms of performance, i.e., computation cost, and space requirement in general. Tables 1 and 2 play a major role in deciding the efficiency of the C-PPA algorithm. Table 1 speaks with respect to space requirement whereas Table 2 shows effectiveness in execution time. The application of C-PPA also performs feature reduction and hence reducing space requirement and computation cost.

Table 2: Comparison of computation time recorded for with and without C-PPA on DT and SVM classifier

Name of dataset	Computation time (seconds)							
	DT		SVM		NB		CNN	
	w/o C-PPA	with C-PPA	w/o C-PPA	with C-PPA	w/o C-PPA	with C-PPA	w/o C-PPA	with C-PPA
Parkinson's	4.70	3.150	5.90	4.00	6.30	5.62	5.66	3.02
Heart disease	0.19	0.03	0.56	0.42	1.25	0.25	0.69	0.03
Kidney	1.58	0.26	2.87	0.14	2.69	1.25	1.58	0.32
Hepatitis	1.61	0.23	3.50	2.10	3.25	2.56	2.36	0.21
Br. Cancer	1.46	0.56	2.87	1.25	2.36	1.25	1.56	0.25
Arrhythmia	2.00	1.35	2.50	1.23	3.69	2.56	1.86	1.12

Scalability includes the efficiency of data mining for increased data size. The feature reduction of C-PPA works for it. C-PPA extracts the input dataset into lower dimensions. More % reduction in features is observed for the datasets having half of the total features near any 2^N. A high level of privacy is achieved by C-PPA as the output of C-PPA is not reversible. The original tuple features are transformed into a smaller number of features which do not have any correlation with the original values. The quality of data is important in the sense of its analysis purpose. If the information is to be mined for an application like health care record, then its accuracy of mining is crucial in decision making. In this regard accuracy, data consistency plays a significant role to assess the quality of data. The accuracy of data mining depends majorly on the information loss after the application of PPDM. From Table 2, it is clear that after the application of C-PPA, the accuracy of classification is either retained or improved. Table 2 shows the computation time elapsed with and without the application of C-PPA. Table 2 observations show that computation time is less with the application of C-PPA as the number of attributes is also reduced. Here, the time complexity of circular convolution using the FFT algorithm, which is the heart of C-PPA, is O (N log N) [47,48].

Table 3 summarizes the observations on accuracy of DT, SVM, NB and CNN classifiers with and without using C-PPA algorithm as well as different feature extraction algorithms like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). It is observed that CNN gives the same or better accuracy as compared to without C-PPA algorithm. The use of C-PPA algorithm shows a 14.15% average gain in accuracy for the CNN classifier. Additionally, the comparison with previous feature reduction techniques shows that the proposed algorithm achieves better accuracy results than existing methods.

From Table 4, it is clear that after the application of C-PPA, the accuracy of classification is either increased or remains close to the original accuracy. For classification, a CNN is applied to the output of C-PPA as CNN achieves higher classification accuracy when compared to other classifier algorithms. Here, C-PPA acts as a feature-reduction technique for classification purposes.

Table 3: Comparison of accuracy for CNN, DT, SVM and NB classifiers with and without C-PPA, PCA, and SVD techniques

Dataset names		Parkinson's disease	Heart	Kidney	Hepatitis	Breast cancer	Arrhythmia	Average Δ
CNN	PCA	81.25	82	86.05	90.58	90.21	90.1	
	SVD	80.54	81.56	88.9	89.65	90.12	91.45	
	w/o C-PPA	79.81	78.56	75.25	80.22	81.25	83.78	
	With C-PPA	81.2	95.23	98.56	96.56	95.68	96.58	14.15
	ΔC-PPA	1.39	16.67	23.31	16.34	14.43	12.8	
DT	PCA	80.2	85	93	84	92.3	92.3	
	SVD	79.5	80.2	90	85.25	90.2	91.5	
	w/o C-PPA	74.6	86.13	95.75	88.63	93.71	94.91	4.88
	With C-PPA	81.08	94.73	97.25	94.69	94.42	95.79	
	ΔC-PPA	8.69	9.98	1.57	6.84	1.29	0.93	
SVM	PCA	72	73.5	95	97.8	91.23	50.5	
	SVD	73.25	73.21	73	95.3	95	90.2	
	w/o C-PPA	74.6	74.91	96.25	98.48	92.7	54.2	
	With C-PPA	74.6	79.71	97	98.69	94.13	54.2	1.49
	ΔC-PPA	0	6.41	0.78	0.21	1.54	0	
NB	PCA	70.25	71.22	94.65	95.23	90.56	50.21	
	SVD	71.56	71.45	72.36	93.56	91	88.69	
	w/o C-PPA	73.25	71.32	95.68	97.85	91.56	50.23	1.53
	With C-PPA	74.2	75.63	95.66	97.68	93.24	52.31	
	ΔC-PPA	0.95	4.31	0.02	0.17	1.68	2.08	

Table 4: Comparison of % accuracy obtained without and with C-PPA for different classifiers

Name of dataset	DT classifier		SVM classifier		CNN classifier		NB classifier	
	w/o C-PPA	With C-PPA	w/o C-PPA	with C-PPA	w/o C-PPA	With C-PPA	w/o C-PPA	with C-PPA
Parkinson's disease	74.06	81.08	74.60	74.60	72.00	82.22	70.32	73.60
Heart	86.13	94.73	74.91	79.71	84.58	95.66	70.25	85.56
Kidney	95.75	97.25	96.25	97.00	80.02	97.55	89.01	95.23
Hepatitis	88.63	94.69	98.48	98.69	88.36	95.63	86.23	92.00
Br. Cancer	93.71	94.92	92.75	94.13	90.25	95.20	88.23	93.25
Arrhythmia	94.91	95.79	54.20	54.20	92.25	94.5	87.12	88.60

In the C-PPA approach as stated earlier, the features can be grouped into more than two groups, processed parallelly then execution time can be reduced further. It is known as overlap-add approach of convolution. Table 5 compares the accuracy without C-PPA with Overlap-add convolution-based PPA algorithm (OAC-PPA). An average gain in accuracy with Overlap-add C-PPA is 9.11, 9.78, 08.15, 12.79, and 10.58 for DT, SVM, NB, and CNN classifiers. Tables 4 and 5 strongly prove the accuracy gain factor of the proposed algorithm by using different classifiers. It is clear from the results that despite of gain in accuracy the important feature is a reduction of features which in turn causes a reduction in processing time and space as well.

Table 6 lists the performance of different classifiers on different parameters by performing the proposed feature reduction technique on different healthcare datasets. Here, the classifiers like SVM, DT, CNN, and NB are analyzed to show the efficiency of the proposed methodology. These classification algorithms are implemented by ourselves, and the comparison is made for different medical datasets. From the comparison, it is shown that the DT classifier displays better results in terms of different parameters like accuracy, precision, recall, and *F*-measure.

Table 5: Comparison of accuracy for with C-PPA, without C-PPA and overlap add convolution-based C-PPA using different classifiers

Datasets	DT classifier		SVM classifier		NB		CNN	
	With OAC	ΔOAC	With OAC	ΔOAC	With OAC	ΔOAC	With OAC	ΔOAC
Parkinson's disease	97.22	30.32	97.22	30.32	95.33	29.65	97.56	33.23
Heart	94.05	09.20	91.08	21.59	92.35	08.56	95.63	25.63
Kidney	99.75	04.18	98.75	02.60	98.65	03.21	99.85	05.65
Hepatitis	91.66	03.42	98.66	00.18	90.26	02.31	98.89	01.25
Breast cancer	98.14	04.73	96.42	04.01	95.65	03.65	98.98	05.65
Arrhythmia	97.56	02.79	54.20	00.00	56.23	01.56	97.68	03.56
Average Δ	**09.11**		**09.78**		**08.15**		**12.49**	

Table 6: Analysis of other parameters on different classifiers

Classifiers		Parkinson's disease	Heart	Kidney	Hepatitis	Breast cancer	Arrhythmia
CNN	Precision	95.23	92.1	96.2	90.56	95.23	93.58
	Recall	94	90	94	88.56	92.78	90
	F-Measure	93	88.6	90.23	85.65	89.56	87.87
SVM	Precision	93.25	88.98	92	92.14	90	56
	Recall	92.3	88	91	90	90.23	61
	F-Measure	91.23	88.20	90	90	90	55
DT	Precision	92.56	87.8	91	91.58	89	61
	Recall	91	87	90	90.7	88.98	60
	F-Measure	90	86.8	87	90	87	59.98
NB	Precision	91	87	90	90.56	88	63
	Recall	89.2	85.36	85	88.56	87	62
	F-Measure	88	84	80	85	85	63

5 Conclusion

Privacy preservation is the most crucial feature expected with any data analytics. Data sharing has become essential for different purposes like healthcare analytics, surveys, surveillance, IoT-based systems for decision-making, and so on. The third-party sharing of data does not guarantee privacy preservation. Existing privacy preservation algorithms are efficient in terms of privacy preservation but lose data analytics accuracy. The proposed algorithm is tested in comparison to the different classifiers like DT, SVM, RF, NB, and LR. The proposed algorithm works well on data classification accuracy with maintaining the secrecy of data. Also, it extracts data in a reduced form to make data time and space efficient. The proposed C-PPA is a lossy algorithm as one cannot recover original data after applying C-PPA. One can share the output of C-PPA on any network, several times without loss of privacy. As discussed above, after applying C-PPA, one can benefit from processing cost, i.e., execution time and space, retaining the privacy of data, and accuracy of data processing. As per the classification, the DT, SVM, CNN, and NB algorithm achieves higher classification accuracy on six utilized datasets.

Funding information: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions: The Authors have contributed in idea inception, system implementation, result collection and analysis.

Conflict of interest: The authors of this publication declare there is no conflict of interest.

Data availability statement: This research uses an online data set which is properly cited in the article and available online.

References

[1] Shen Y, Han J, Shao H. Research on privacy-preserving technology of data mining. 2009 Second International Conference on Intelligent Computation Technology and Automation; 2009. p. 612–4. doi: 10.1109/ICICTA.2009.382.
[2] Girka A, Terziyan V, Gavriushenko M, Gontarenko A. Anonymization as homeomorphic data space transformation for privacy-preserving deep learning. Procedia Comput Sci. 2021;180:867–76. ISSN 1877-0509.
[3] Aldeen YAAS, Salleh M, Razzaque MA. A comprehensive review on privacy preserving data mining. Springer Plus. 2015;4:694. doi: 10.1186/s40064-015-1481-x.
[4] Fouad H, Hassanein AS, Soliman AM, Al-Feel H. Analyzing patient health information based on IoT sensor with AI for improving patient assistance in the future direction. Measurement. 2020;159:107757. ISSN 0263-2241. doi: 10.1016/j.measurement.2020.107757.
[5] Domadiyaa N, Rao UP. Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. Procedia Comput Sci. 2019;148:303–12.
[6] Zorarpac E, Özel SA. Privacy preserving classification over differentially private data. WIRES Data mining and knowledge discovery. United States: John Wiley & Sons Inc.; 2020. doi: 10.1002/widm.1399.
[7] Han J, Kamber M. Data mining: Concepts and techniques. 3rd edn. USA: Morgan Kaufmann Publishers; 2006.
[8] Ge W, Wang W, Li X, Shi B. A privacy-preserving classification mining algorithm. In: Ho TB, Cheung D, Liu H, editors. Advances in Knowledge Discovery and Data Mining. PAKDD 2005. Lecture Notes in Computer Science, Vol. 3518. Berlin: Springer, Heidelberg; 2005. https://doi.org/10.1007/11430919_32.
[9] Fatehi N, Shahhoseini HS, Wei J, Chang C-T. An automata algorithm for generating trusted graphs in online social networks. Appl Soft Comput. 2022;118:108475. ISSN 1568-4946. doi: 10.1016/j.asoc.2022.108475.
[10] Zhou X, Xu K, Wang N, Jiao J, Dong N, Han M, et al. A secure and privacy-preserving machine learning model sharing scheme for edge-enabled IoT. IEEE Access. 2021;9:17256–65. doi: 10.1109/ACCESS.2021.3051945.
[11] Niu C, Wu F, Tang S, Ma S, Chen G. Toward verifiable and privacy preserving machine learning prediction. IEEE Trans Dependable Secure Comput. 2022;19:1703–21. doi: 10.1109/TDSC.2020.3035591.
[12] Mohassel P, Zhang Y. SecureML: A system for scalable privacy-preserving machine learning. 2017 IEEE Symposium on Security and Privacy (SP); 2017. p. 19–38. doi: 10.1109/SP.2017.12.
[13] Shokri R, Shmatikov V. Privacy-preserving deep learning. 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton); 2015. p. 909–10. doi: ALLERTON.2015.7447103">10.1109/ALLERTON.2015.7447103.
[14] Aggarwal CC. On k-anonymity and the curse of dimensionality. VLDB Conference; 2005.
[15] Sweeney L. K-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst. 2002;10(5):557–70. (October 2002). doi: 10.1142/S0218488502001648.
[16] Basu A, Nakamura T, Hidano S, Kiyomoto S. K-anonymity: Risks and the Reality. Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA - Volume 01 (TRUSTCOM '15). USA: IEEE Computer Society; 2015. p. 983–9. doi: 10.1109/Trustcom.2015.473.
[17] Rafiei M, van der Aalst WMP. Group-based privacy preservation techniques for process mining. Data Knowl Eng. 2021;134:101908.
[18] Lee J-S, Jun S-P. Privacy-preserving data mining for open government data from heterogeneous sources. Gov Inf Q. 2021;38:101544.
[19] Cunha M, Mendes R, Vilela JP. A survey of privacy-preserving mechanisms for heterogeneous data types. Comput Sci Rev. 2021;41:100403.
[20] Matatov N, Rokach L, Maimon O. Privacy-preserving data mining: A feature set partitioning approach. Inf Sci. 2010;180:2696–720; Iverson LI, Iverson SD, Snyder SH. Handbook of psychopharmacology. Vol II. New York: Plenum Press; 2020. p. 99–115.
[21] Gong M, Xie Y, Pan K, Feng K, Qin AK. A survey on differentially private machine learning [Review Article]. IEEE Comput Intell Mag. 2020;15(2):49–64. doi: 10.1109/MCI.2020.2976185.
[22] Hababeh I, Gharaibeh A, Nofal S, Khalil I. An integrated methodology for big data classification and security for improving cloud systems data mobility. IEEE Access. 2019;7:9153–63. doi: 10.1109/ACCESS.2018.2890099.
[23] Samaraweera GD, Chang JM. Security and privacy implications on database systems in big data era: A survey. IEEE Trans Knowl Data Eng. 1 Jan. 2021;33(1):239–58. doi: 10.1109/TKDE.2019.2929794.
[24] Jing X, Yan Z, Pedrycz W. Security data collection and data analytics in the internet: A survey. IEEE Commun Surv Tutorials. Firstquarter 2019;21(1):586–618. doi: 10.1109/COMST.2018.2863942.

[25] Wang X, Ishii H, Du L, Cheng P, Chen J. Differential Privacy-preserving Distributed Machine Learning. 2019 IEEE 58th Conference on Decision and Control (CDC); 2019. p. 7339–44. doi: 10.1109/CDC40024.2019.9029938.
[26] Jiang B, Li J, Yue G, Song H. Differential privacy for industrial internet of things: Opportunities, applications, and challenges. IEEE Internet Things J. 2021;8(13):10430–51. doi: 10.1109/JIOT.2021.3057419.
[27] Almagrabi AO, Bashir AK. A classification-based privacy-preserving decision-making for secure data sharing in internet of things assisted applications. Digital Commun Netw. 2021;8:436–45. ISSN 2352-8648.
[28] Sharma P, Borah MD, Namasudra S. Improving security of medical big data by using Blockchain technology. Comput Electr Eng. 2021;96(Part A):107529. ISSN 0045-7906. doi: 10.1016/j.compeleceng.2021.107529.
[29] Zhang T, Zhu T, Xiong P, Huo H, Tari Z, Zhou W. Correlated differential privacy: Feature selection in machine learning. IEEE Trans Industrial Inf. March 2020;16(3):2115–24. doi: 10.1109/TII.2019.2936825.
[30] Mathew G, Obradovic Z. Poster: Auto-reduction of features for containing communication costs in a distributed privacy-preserving clinical decision support system. 2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS); 2013. p. 1. doi: 10.1109/ICCABS.2013.6629206.
[31] Skubalska-Rafajłowicz E. Spatially-organized random projections of images for dimensionality reduction and privacy-preserving classification. 2017 10th International Workshop on Multidimensional (nD) Systems (nDS); 2017. p. 1–5. doi: 10.1109/NDS.2017.8070627; Hecker AL. Nutrition and physical performance. In RH Strauss, editor. Drugs & performance in sport (2nd edn). Philadelphia: WB Saunders; 2018. p. 23–40.
[32] Nazir S, Patel S, Patel D. Autoencoder based anomaly detection for SCADA networks. Int J Artif Intell Mach Learn. 2021;11(2):83–99. doi: 10.4018/IJAIML.20210701.oa6.
[33] Ahmed N, Deng Z, Memon I, Hassan F, Mohammadani KH, Iqbal R. A survey on location privacy attacks and prevention deployed with IoT in vehicular networks. Wirel Commun Mob Comput. 2022 Apr 26;2022;2022:1–15. doi: 10.1155/2022/6503299.
[34] Junejo MH, Ab Rahman AA, Shaikh RA, Yusof KM, Kumar D, Memon I. Lightweight trust model with machine learning scheme for secure privacy in VANET. Procedia Comput Sci. 2021 Jan 1;194:45–59.
[35] Memon I, Shaikh RA, Hasan MK, Hassan R, Haq AU, Zainol KA. Protect mobile travelers information in sensitive region based on fuzzy logic in IoT technology. Secur Commun Netw. 2020;2020:1–12. doi: 10.1155/2020/8897098.
[36] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences. 2013 Apr 9;110(15):5802–5.
[37] Jia J, Gong NZ. {AttriGuard}: A practical defense against attribute inference attacks via adversarial machine learning. 27th USENIX Security Symposium (USENIX Security 18). 2018;513–29. https://www.usenix.org/conference/usenixsecurity18/presentation/jia-jinyuan.
[38] Almaiah MA, Ali A, Hajjej F, Pasha MF, Alohali MA. A lightweight hybrid deep learning privacy preserving model for FC-based industrial internet of medical things. Sensors. 2022 Mar 9;22(6):2112.
[39] Duan J, Zhou J, Li Y, Huang C. Privacy-preserving and verifiable deep learning inference based on secret sharing. Neurocomputing. 2022 Apr 28;483:221–34.
[40] Alkhelaiwi M, Boulila W, Ahmad J, Koubaa A, Driss M. An efficient approach based on privacy-preserving deep learning for satellite image classification. Remote Sens. 2021 Jun 6;13(11):2221.
[41] Proakis JG, Manolakis. DK. Digital signal processing: Principles, algorithms, and applications. 3rd Edn. New Jersy: Pearson Publications; 1996.
[42] Tajanpure R, Muddana A. Circular convolution-based feature extraction algorithm for classification of high-dimensional datasets. J Intell Syst. 2021;30(1):1026–39. doi: 10.1515/jisys-2020-0064.
[43] Bertino E, Fovino IN, Provenza LP. A Framework for Evaluating Privacy Preserving Data Mining Algorithms*. Data Min Knowl Disc. 2005;11:121–54. doi: 10.1007/s10618-005-0006-6.
[44] Bertino E, Lin D, Jiang W. A survey of quantification of privacy preserving data mining algorithms, privacy-preserving data mining: Models and algorithms. US: Springer. p. 183–205ISBN 978-0-387-70992-5. doi: 10.1007/978-0-387-70992-5_8.
[45] Qi X, Zong M. An overview of privacy preserving data mining. 2011 International Conference on Environmental Science and Engineering (ICESE 2011). Vol. 12; 2012. p. 1341–7. Procedia Environmental Sciences.
[46] Oppenheim AV, Schafer RW. Digital signal processing. 1st edn. The University of Michigan, Pearson; Jan 12 1975.
[47] Tajanpure R, Muddana A. Overlapped circular convolution based feature extraction algorithm for classification of high dimensional datasets. In Singh M, Tyagi V, Gupta PK, Flusser J, Ören T, Sonawane VR, editors. Advances in Computing and Data Sciences. ICACDS 2021. Communications in Computer and Information Science. Vol. 1440, Cham: Springer; 2021. doi: 10.1007/978-3-030-81462-5_20; Aldeen YAAS, Salleh M, Razzaque MA. A comprehensive review on privacy preserving data mining. Springer Plus 2015;4:694. doi: 10.1186/s40064-015-1481-x.
[48] Liu B, Ding M, Shaham S, Rahayu W, Farokhi F, Lin Z. When machine learning meets privacy a survey and outlook. ACM Comput Surv. 2021;54. doi: 10.1145/3436755 (A through study on privacy preservation and machine learning approaches is done by author Liu et al. Author surveyed on private machine learning(ML), ML with privacy protection and privacy attacks on ML.).