



STIF: Intuitionistic fuzzy Gaussian membership function with statistical transformation weight of evidence and information value for private information preservation

G. Sathish Kumar¹ · K. Premalatha²

Accepted: 10 April 2023 / Published online: 21 April 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Data sharing to the multiple organizations are essential for analysis in many situations. The shared data contains the individual's private and sensitive information and results in privacy breach. To overcome the privacy challenges, privacy preserving data mining (PPDM) has progressed as a solution. This work addresses the problem of PPDM by proposing statistical transformation with intuitionistic fuzzy (STIF) algorithm for data perturbation. The STIF algorithm contains statistical methods weight of evidence, information value and intuitionistic fuzzy Gaussian membership function. The STIF algorithm is applied on three benchmark datasets adult income, bank marketing and lung cancer. The classifier models decision tree, random forest, extreme gradient boost and support vector machines are used for accuracy and performance analysis. The results show that the STIF algorithm achieves 99% of accuracy for adult income dataset and 100% accuracy for both bank marketing and lung cancer datasets. Further, the results highlights that the STIF algorithm outperforms in data perturbation capacity and privacy preserving capacity than the state-of-art algorithms without any information loss on both numerical and categorical data.

Keywords Classifiers · Information value · Intuitionistic fuzzy · Privacy preserving data mining · Weight of evidence

✉ G. Sathish Kumar
saathhish@gmail.com

K. Premalatha
kpl_barath@yahoo.co.in

¹ Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India

² Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamil Nadu, India

1 Introduction

Huge amount of data archives are offered by the field of business, education, industry, science and technology. Those data can afford a rich source for knowledge discovery and decision support in data mining. Data mining is a multidisciplinary approach for data analyse and ultimately use those data [1]. If the data has been given for mining, there is every chance of security and privacy breach. Thus, a method is essential to disclose the private and sensitive data for analysis without any privacy breach and balancing the data utility thus making the significant area of research. Privacy Preserving Data Mining (PPDM) results in securing the user's private data without any data loss. PPDM manages the numerical data by focusing on the statistical measures between the actual and perturbed data [2].

The PPDM technique discourses the concern about privacy preservation by data perturbation and distribution reconstruction at aggregate level to perform data mining. The distribution reconstruction approach leads to information loss at the acceptable rate in real-world circumstances by using expectation maximization algorithms [3]. In PPDM, fuzzy logic technique has been applied for data perturbation. Fuzzy logic is a form of many-valued logic to deal the degree of truth and expressed in any real number between 0 to 1. There are several extensions has been developed since the introduction of fuzzy set. The intuitionistic fuzzy set plays the major role among those developments. Interval-valued intuitionistic fuzzy sets (IVIFSs) have evolved to evaluate uncertainty and risk in the data. These IVIFSs are further applied to the loss functions of decision-theoretic rough sets (DTRSs). The DTRSs is the optimization approach for IVIFSs three-way decisions [4]. The three-way decision is used to classify the data into three parts based on positive, negative and boundary regions. In the knowledge discovery process, optimal scale combination selection and attribute reduction are the key concern. To overcome this issue, the three-way decision is applied to perform scale selection and attribute reduction simultaneously [5]. In addition, the three-way decision is evaluated on MovieLens-100 K, MovieLens-1 M and MovieLens-10 M datasets to find the dynamic variations of user's data and performance using the recommender systems [6]. Intuitionistic Fuzzy Set (IFS) is the generalization of the notion of fuzzy set (FS). The FS and IFS have the form $\{\langle x, \mu_A(x) \rangle | x \in E\}$ and $\{\langle x, \mu_A(x), 1 - \mu_A(x) \rangle | x \in E\}$ respectively. In addition the IFS have degree of indeterminacy or uncertainty, which is expressed as $\pi(x) = 1 - \mu(x) - \nu(x)$ [7, 8]. This uncertainty is called as the hesitation index. At times, it is tedious to determine the membership degree of an element into the fixed set. IFS provide the solution using the hesitation index which is characterized by the membership and non-membership function [9].

In this article, Intuitionistic fuzzy technique has been employed to perturb the data. In addition, the statistical transformation methods like weight of evidence (WOE) and information value (IV) are used for data perturbation. WOE and IV are logistic regression models to solve the classification problems [10]. Initially, the user's private and sensitive data is perturbed with WOE and IV. The resultant value is given as the input for intuitionistic fuzzy for further perturbation

to enhance the data privacy. Consider the pharmaceutical data of a customer. It contains the customer's private information such as name, age, sex, address and even the phone number. Using the pharmaceutical data it is very easy to predict the customer's sensitive data like disease and the nature of disease. If these data are given to the external agencies for analysis, they can extract the private and sensitive data about the individual and can use it for illegal activities. To avoid those kinds of security breaches, the data should be perturbed before it get into the hands of third parties. Here, the proposed algorithm solves the issue of privacy breach. The proposed algorithm can be applied for both numerical and categorical data for perturbation. After perturbation, the perturbed data can be given for the external parties for analysis. The external parties will not be able to extract private and sensitive data about an individual. The knowledge extracted before and after perturbation remains the same, thus preserves the data utility in line with data privacy.

The rest of the paper is structured as follows: Motivation of the research work is exposed in Sect. 2, related work in the area of PPDM is discussed in Sect. 3, Sect. 4 illustrates the analysis of research gap, Sect. 5 represents the proposed methodology to carry out the process of privacy preservation, implementation results were highlighted in Sect. 6 and finally the conclusion is given in Sect. 7.

2 Motivation for the research

Mainly in the Covid-19 pandemic, all education and business have been done over the internet. A vast amount of data has been deposited into the repositories during the pandemic. Those data contain numerous private and sensitive information about the individuals. This creates a major concern for people about their privacy. Privacy is a basic and essential right for every human being in this internet world. The privacy breach not only affects the individual, but also creates a serious threat to the nation's security.

This privacy concern has sown the seed for this research work. Many related works have been studied and analysed to understand the usage of privacy preserving algorithms. Many existing algorithms guarantee the user's privacy but they lack data utility and vice-versa. In this research work, intuitionistic fuzzy Gaussian membership function and statistical transformation methods are considered for developing the novel statistical transformation with an intuitionistic fuzzy algorithm to preserve the user's privacy along with data utility. The STIF algorithm has been applied to both numerical and categorical data for performing data perturbation. The STIF applied privacy preservation is considered as a significant feature for the effective use of an immense volume of data and can share personal data for analysis. After the data perturbation using STIF algorithm, the commercial third parties cannot extract any private and sensitive data during analysis from the shared data. This is the real success of this research work.

3 Related work

Numerous methods have been proposed to mask the private and sensitive data. There are couple of methods used widely for preserving the data privacy. Randomization method and cryptographic method are employed on centralized data and distributed data respectively. The randomized function has been applied by the data miner to recognize the aggregated property in randomization method [11]. A frequent mining model based on the deep learning network approach and paillier encryption technique has been employed to reduce unauthorized data access, computation complexity and provides the data privacy [12]. This approach does not care for the data utility if the expectation results in more precise mining.

In cryptographic method, the encrypted data has been shared for analysis to the external parties. The data miner can only decrypt the aggregate result from the encrypted data rather than decrypting the individual result. This results in the false prediction of data. The cryptographic approach yields the greater security but results in aggregate information loss. Secure multiparty computation (SMC) is a best example for such approach [13]. However, the SMC has high computation costs on the customer side. C4.5 algorithm along with the weighted average protocol and secure set intersection protocol has been used as the remedy for computation cost problems and to yield privacy [14]. The properties of homomorphic encryption and custom secure communication protocol are used to securely share and mine the association rules from several parties with the assertion of privacy [15].

Person-specific and field structured data from the hospital or a bank contains the private data of an individual. l –diversity and k –Anonymity model has been applied on person-specific and field structured data to protect the data privacy. Enhanced diversity model is applied to the foursquare dataset and MIMIC-III dataset to perturb the data by adding or deleting the spatial temporal values by satisfying ϵ -privacy [16]. The k –anonymity protects the information if release contains the $k - 1$ individual's whose data also present in the release. It is difficult to identify the actual data from the $k - 1$ records. But the k -anonymity results in temporal attack due to dynamic data collection [17]. Light weight edge-based classifier method is applied over the encrypted data using edge computing which has paved the way for data security and user query privacy [18]. This k –Anonymity is often considered only for the attributes containing the information for differentiating a particular sample from the simple set. A fuzzy variant of k –member clustering is offered to increase the quality of data summarization with k –anonymity. Fuzzy membership degree has been considered for k –member cluster for estimating the cluster distance [19]. The best distance measure has not been predicted by the fuzzy variant of k –member clustering.

Fuzzy set theory for data classification has been used to preserve the privacy. Data points have been transformed to fuzzy attributes by maintaining the similarity between the data points. Naïve Bayesian and ID3 classification algorithms are used to measure the classification accuracy. Association rule mining has not been

applied and not tested for various other classification algorithms [20]. Randomization process for data perturbation has been generated from the fuzzy system based on level of privacy and various inputs [21]. Privacy preserving rewiring algorithm uses the fuzzy set technique for the process of data fuzzification. This fuzzification has been applied to social network data and quantitative data for privacy preservation [22]. The privacy preservation of quantitative data by hiding the cluster of exciting patterns holds the sensitive information. The perturbation on that sensitive knowledge has the minimum side effects like lost rules and ghost rules. Fuzzy association rule along with the famous particle swarm optimization algorithm has applied to perturb the quantitative data with least number of modifications [23].

Fuzzy set and apriori mining algorithms are integrated to discover the fuzzy association rule to perform data hiding. Quantitative database have been used to formulate the fuzzy association rule hiding algorithm for hiding rules [24]. Classification algorithms use the weight of evidence as a tool for data perturbation. The perturbed data can be used for all the supervised learning problems [25]. The weight of evidence can be applied with fuzzy set to perform data transformation to achieve the privacy.

Rough set theory on intuitionistic fuzzy yields the data reduction and hiding the sensitive rules. The sensitive fuzzy association rule has been hidden by the sensitive rule hiding approach. The fuzzified system has been built by using the trapezoidal and triangular membership functions [26]. Hesitant function has been used to choose the best membership degree of an attribute to a set contains the different values. The hesitant function will return the subset value between 0 to 1 [27]. Fuzzy clustering method has been used to provide the intuitionistic fuzzy partition for the application of privacy. This approach has been applied to execute the many fuzzy functions like fuzzy c -means and entropy based fuzzy c -means for privacy preservation [28]. The results of fuzzy optimization and intuitionistic fuzzy optimization methods are compared based on exponential and uniform lead time demand and found that intuitionistic fuzzy model offers the better result than the ordinary fuzzy optimization models [29]. The classifiers decision tree, random forest, support vector machines and extreme gradient boost are used for analysis [30].

Several types of intuitionistic fuzzification functions such as triangular, trapezoidal, Gaussian, bell-shaped, sigmoidal, S-shaped, Z-shaped functions has been examined to address the issue of uncertainty and to model real world situations in intuitionistic fuzzy environment. Gaussian membership function can be effectively applied for the numerical values to perform the fuzzification process [31].

4 Research gap analysis

From the literature review, many studies have discussed the protection of privacy breaches [2, 3] and balancing the data utility to some extent. Cryptographic approach [13, 14], l -diversity and k -anonymity [9, 17] are mostly used for data perturbation. These approaches are efficient in securing data privacy but result in higher computation costs and temporal attacks [17] due to dynamic data pooling in the

repositories. Also, data classification techniques are applied using the fuzzy set for privacy preservation [20, 24, 26]. Most of the existing data perturbation algorithms are been applied only to the numerical data and have not been applied to categorical data in the transactional datasets [30]. So, a better privacy preservation algorithm could improve data privacy on both numerical and categorical data without any information loss.

4.1 Objective of the research

Based on the literature review and research gap found in the studies, the following objectives are considered for this research work.

- To develop a novel algorithm for privacy preservation without negotiating the data utility.
- To apply the proposed algorithm to both numerical and categorical data.
- To analyse the proposed algorithm by measuring the performance metrics for the detected results.
- To evaluate the performance of the proposed algorithm by comparing the performance metrics of algorithms from earlier studies.

As many essential abbreviations are used in this research paper, it is listed below for easy and better understanding.

AUC	Area Under The Curve
DPC	Data Perturbation Capacity
DPS	Differential Privacy Scheme using Gaussian Mechanism
DRC	Data Retrieval Capacity
DT	Decision Tree
DTRSs	Decision Theoretic Rough Sets
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FS	Fuzzy Set
GMF	Gaussian Membership Function
HEA	Homomorphic Encryption using ElGamal Algorithm
IFS	Intuitionistic Fuzzy Set
ITMF	Intuitionistic Fuzzy Triangular Membership Function
ITrMF	Intuitionistic Fuzzy Trapezoidal Membership Function
IV	Information Value
IVIFSs	Interval-Valued Intuitionistic Fuzzy Sets
PPC	Privacy Preserving Capacity
PPDM	Privacy Preserving Data Mining
RF	Random Forest
ROC	Receiver Operating Characteristic
SMC	Secure Multiparty Computation

STIF	Statistical Transformation With Intuitionistic Fuzzy
SVM	Support Vector Machines
TMF	Triangular Membership Function
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TrMF	Trapezoidal Membership Function
UCI	University Of California, Irvine
WOE	Weight Of Evidence
XGB	Extreme Gradient Boost

5 Proposed methodology

In this article, STIF algorithm has been proposed to perform the data perturbation. The STIF algorithm comprises the statistical transformation method such as WOE and IV along with Intuitionistic fuzzy approach for privacy preservation of user's data. The STIF algorithm has been employed on three standard datasets from University of California, Irvine (UCI) machine learning repository which includes adult Income, bank marketing and lung cancer datasets for analysis. The STIF algorithm has been applied to both the numerical data and categorical data for perturbation. The data perturbation using STIF contains two phases. In phase-1, the statistical transformation methods WOE and IV are applied for initial data perturbation. To enhance the privacy preservation capacity, the intuitionistic fuzzy approach has been applied on the initially perturbed data in phase-2. Figure 1 illustrates the data perturbation process using the proposed STIF algorithm.

5.1 Phase 1—Statistical transformation using weight of evidence and information value

The first step in data perturbation is to find the WOE and IV for the original data. The WOE and IV is the highly powerful statistical techniques evolved from logistic regression model. These techniques can be applied to both numerical and categorical data. The formula to calculate WOE and IV is given in the Eqs. 1 and 2.

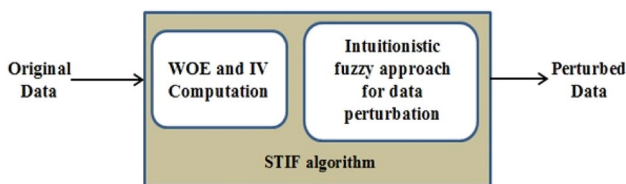


Fig. 1 Data perturbation using STIF algorithm

$$\text{WOE} = \ln(\text{instance\%} / \text{non_instance\%}) \quad (1)$$

$$\text{IV} = \sum (\text{instance\%} - \text{non_instance\%}) * \text{WOE} \quad (2)$$

Here, instance refers to the attributes with positive results and non_instance refers to the attributes with negative results. These results are based on the class label in the dataset.

5.1.1 Steps for computing the values of WOE and IV

Using the Eqs. 1 and 2, the value for WOE and IV can be computed by the following sequence of steps for continuous attribute and categorical attribute.

5.1.1.1 WOE and IV for continuous attribute

- Step 1: For Continuous attribute, divide the data into 'X' parts (bins).
- Step 2: Calculate the value for instance and non_instance for all the bins.
- Step 3: Compute the %ofinstance and %ofnon_instance for all the bin values.
- Step 4: Perform division operation between %ofinstance and %ofnon_instance.
- Step 5: Find the WOE by computing natural log (ln) for the resultant value from step 4.
- Step 6: Calculate the difference between %ofinstance and %ofnon_instance.
- Step 7: Find the IV by computing the product between WOE value and the resultant value from step 6.

Consider duration attribute from bank marketing dataset. The duration attribute is divided into 9 bins to compute the value of instance and non_instance based on the class label. Further, the percentage of instance and non_instance are calculated to find WOE and IV. Table 1 represents the statistical transformation value using WOE and IV for duration (numerical) attribute.

5.1.1.2 WOE and IV for categorical attribute For categorical attribute, dividing the data into X parts (bins) is not required. But, need to perform all the operations from step-2 to step-7 as mentioned in the Sect. 5.1.1.1 for computing WOE and IV for categorical values.

Consider occupation attribute from adult income dataset. There are 14 occupation categories found in the occupation attribute. Based on the class label, the value for instance and non_instance has been calculated. In turn, the percentage of instance and non_instance has computed to find the resultant value of WOE and IV for categorical values. Table 2 depicts the statistical transformation value using WOE and IV for occupation (categorical) attribute.

Table 1 WOE and IV computation for numerical attribute

Range	Bins	Count	Instance	Non_ Instance	% of Instance	% of Non_ Instance	WOE	IV
0–500	1	36,260	2542	33,718	7.01	92.99	−2.59	222.69
501–1000	2	3972	1532	2440	38.57	61.43	−0.47	10.74
1001–1500	3	750	437	313	58.27	41.73	0.33	5.46
1501–2000	4	146	91	55	62.33	37.67	0.51	12.58
2001–2500	5	36	25	11	69.44	30.56	0.82	31.88
2501–3000	6	9	7	2	77.78	22.22	1.25	69.45
3001–3500	7	9	3	6	33.33	66.67	−0.69	23.01
3501–4000	8	5	3	2	60	40	0.41	8.2
> 4000	9	3	2	1	66.66	33.34	0.69	22.91
Total			4642	36,548				

Table 2 WOE and IV computation for categorical attribute

Occupation	Count	Instance	Non_ Instance	% of Instance	% of Non_ Instance	WOE	IV
Administrator-clerical	3770	3263	507	86.55	13.45	1.86	135.97
Armed Forces	9	8	1	88.89	11.11	2.08	161.78
Craft-Repair	4099	3170	929	77.34	22.66	1.23	67.25
Executive-Manager	4066	2098	1968	51.60	48.40	0.07	0.22
Farming-Fishing	994	879	115	88.43	11.57	2.03	156.03
Handlers-Cleaners	1370	1284	86	93.72	6.28	2.7	236.10
Machine-opt-inspect	2002	1752	250	87.51	12.49	1.95	146.30
Other Service	3295	3158	137	95.84	4.16	3.14	287.89
Private house servant	149	148	1	99.33	0.67	4.99	492.30
Prof.-specialty	4140	2281	1859	55.10	44.90	0.2	2.04
Protective Service	649	438	211	67.49	32.51	0.73	25.53
Sales	3650	2667	983	73.07	26.93	0.99	45.68
Tech Support	928	645	283	69.50	30.50	0.82	31.99
Transport moving	1597	1277	320	79.96	20.04	1.38	82.70
Total		23,068	7650				

As shown in Tables 1 and 2, the WOE and IV values can be calculated for all the quasi-identifiers of numerical and categorical types in adult income, bank marketing and lung cancer datasets to perform the initial data perturbation. The initially perturbed value is passed as the input for intuitionistic fuzzy logic approach to perform second level of perturbation to enhance the privacy preserving capacity in phase 2.

Table 3 Initial data perturbation using IV for adult income dataset

Operation	Age	Education	Occupation
Original value	39	9th	Sales
After calculating IV	7.17	24.28	46.68

Table 3 shows the actual value and the perturbed value using IV for age, education and occupation attribute form adult income dataset. This resultant IV value should be given as the input for intuitionistic fuzzy as the second phase of data perturbation.

5.2 Preliminaries for Phase 2 of data perturbation

In the preliminary phase, the original data is perturbed using the FS and the IFS by applying the triangular, trapezoidal and Gaussian membership functions.

The below sub-sections presents the experimentations done for all the three membership functions on FS and IFS using the adult income, bank marketing and lung cancer datasets. It is more evident from the result, that the IFS with Gaussian membership function provides the highest privacy preserving capacity and data utility than other functions. The data perturbation result by FS and IFS has been done and it is given in the Table 4.

5.2.1 Data perturbation using FS approach

The FS is a mapping of set of real numbers (A_i) onto membership values (x_i) that lies in the range of $[0, 1]$. FS can be represented by (A, x) , where A is a set and x is the membership function denoted by $x : A \rightarrow [0, 1]$, the function $x = \mu_B$ is called as the membership function of fuzzy set $B = (A, x)$. In general, the FS can be formulated as shown in Eq. 3.

Table 4 Data perturbation using fuzzy set for age attribute from adult income dataset after calculating IV by applying TMF, TrMF and GMF

Operation	Age attribute				
	39	50	38	53	28
Original Values					
After calculating IV	7.17	7.28	1.90	2.68	7.11
TMF (Eq. 4)	0.991	0.777	0.973	0.719	0.509
TrMF (Eq. 5)	1.00	0.952	1.00	0.880	1.00
GMF (Eq. 6)	0.999	0.704	0.998	0.581	0.740
Perturbed data using TMF and Eq. (15)	0.991	0.777	0.973	0.719	0.509
Perturbed data using TrMF and Eq. (15)	1.00	0.952	1.00	0.880	1.00
Perturbed data using GMF and Eq. (15)	0.999	0.704	0.998	0.581	0.740

$$A = \{ \langle x, \mu_A(x) \rangle | x \in X \} \quad (3)$$

where $\mu_A(x)$ defines the membership degree of $x \in X$.

5.2.1.1 Triangular membership function (TMF) The TMF contains three arguments, lower limit ‘a’, upper limit ‘c’ and a value ‘b’, where $a \leq b \leq c$. Here, the value of a and c gives the feet of the triangle and the argument b gives the peak of the triangle [31]. TMF is given in the Eq. 4.

$$\mu_A(x) = \begin{cases} 0; & x \leq a \\ \left(\frac{x-a}{b-a} \right); & a \leq x \leq b \\ \left(\frac{c-x}{c-b} \right); & b \leq x \leq c \\ 0; & x \geq c \end{cases} \quad (4)$$

Here, x indicates the membership value.

5.2.1.2 Trapezoidal membership function (TrMF) The TrMF is called as the truncated triangle with flatted top end. TrMF has the four arguments, lower limit ‘a’, upper limit ‘d’, lower support limit ‘b’ and upper support limit ‘c’ where $a \leq b \leq c \leq d$. Here, the argument a and d indicates the trapezium feet and the argument b and c indicates the flatted top end of trapezium [21, 31]. TrMF is given in the Eq. 5.

$$\mu_A(x) = \begin{cases} 0; & x \leq a \\ \left(\frac{x-a}{b-a} \right); & a \leq x \leq b \\ 1; & b \leq x \leq c \\ \left(\frac{x-d}{c-d} \right); & c \leq x \leq d \\ 0; & x \geq d \end{cases} \quad (5)$$

Here, x denotes the membership value.

5.2.1.3 Gaussian membership function (GMF) The Gaussian function is used to represent the vague and linguistic terms of the membership function [8, 31]. The GMF is given in the Eq. 6.

$$\mu_A(x) = \exp(-(x - m)^2 / (2(k)^2)) \quad (6)$$

where m represents the centre, k represents the width of the fuzzy set A and x denotes the membership value.

5.2.2 Data perturbation using IFS approach

Assume the universal set as X . An IFS ‘ D ’ in X can be defined in the form of Eq. 7.

$$A = \{ \langle x, \mu_A(x), \nu_A(x) \rangle | x \in X \} \quad (7)$$

Here $\mu_A(x) \rightarrow [0, 1]$ and $\nu_A(x) \rightarrow [0, 1]$ where, $\mu_A(x)$ and $\nu_A(x)$ defines the membership degree and non membership degree of $x \in X$ respectively.

For intuitionistic fuzzy subset A on the set X , we have hesitancy index $\pi_A(x)$, where x in A . Hesitancy index is determined by

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \quad (8)$$

where $0 \leq \pi_A(x) \leq 1$. $\pi_A(x)$ is the degree of indeterminacy of $x \in X$ to the IFS ‘ A ’.

In addition to the membership function discussed in Sect. 5.2.1, the IFS has the non-membership function and hesitation index.

5.2.2.1 Intuitionistic fuzzy triangular membership function (ITMF) The ITMF contains both the membership function and non-membership function as given in the Eqs. 9 and 10.

$$\mu_A(x) = \begin{cases} 0; & x \leq a \\ \left(\frac{x-a}{b-a} \right); & a \leq x \leq b \\ \left(\frac{c-x}{c-b} \right); & b \leq x \leq c \\ 0; & x \geq c \end{cases} \quad (9)$$

$$\nu_A(x) = 1 - \begin{cases} 0; & x \leq a \\ \left(\frac{x-a}{b-a} \right); & a \leq x \leq b \\ \left(\frac{c-x}{c-b} \right); & b \leq x \leq c \\ 0; & x \geq c \end{cases} \quad (10)$$

The hesitation index

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \quad (11)$$

5.2.2.2 Intuitionistic fuzzy trapezoidal membership function (ITrMF) The membership and non-membership function for the ITrMF along with the hesitation index is given in the Eqs. 12, 13 and 14.

$$\mu_A(x) = \begin{cases} 0; & x \leq a \\ \left(\frac{x-a}{b-a}\right); & a \leq x \leq b \\ 1; & b \leq x \leq c \\ \left(\frac{x-d}{c-d}\right); & c \leq x \leq d \\ 0; & x \geq d \end{cases} \quad (12)$$

$$\nu_A(x) = 1 - \begin{cases} 0; & x \leq a \\ \left(\frac{x-a}{b-a}\right); & a \leq x \leq b \\ 1; & b \leq x \leq c \\ \left(\frac{x-d}{c-d}\right); & c \leq x \leq d \\ 0; & x \geq d \end{cases} \quad (13)$$

The hesitation index is,

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \quad (14)$$

5.2.3 Euclidean n– norm equation

On the n– dimensional space, Euclidean space R^n for the length of vector S is captured by the formulae as given in Eq. 15.

$$|S|_2 = \sqrt{(s_1^2 + s_2^2 \dots + s_n^2)} \quad \text{Where, } S = (S_1, S_2, \dots, S_n) \quad (15)$$

The quasi-identifiers from the adult income, bank marketing and lung cancer dataset is applied on the Eqs. (4, 5, 6, 9, 10, 12, 13) to get the membership value

Table 5 Data perturbation using intuitionistic fuzzy set for age attribute from adult income dataset after calculating IV by applying ITMF and ITrMF

Operation	Age attribute				
	39	50	38	53	28
Original Values					
After calculating IV	7.17	7.28	1.90	2.68	7.11
Membership degree for ITMF (Eq. 9)	0.991	0.777	0.973	0.719	0.509
Non membership degree for ITMF (Eq. 10)	− 0.091	0.122	− 0.073	0.180	0.390
Membership degree for ITrMF (Eq. 12)	1.00	0.952	1.00	0.880	1.00
Non membership degree for ITrMF (Eq. 13)	− 0.10	− 0.05	− 0.10	0.02	− 0.10
Hesitation index—equation (Eq. 8)	0.1	0.1	0.1	0.1	0.1
Perturbed data using ITMF and Eq. 15	1.00	0.79	0.98	0.74	0.65
Perturbed data using ITrMF and Eq. 15	1.00	0.960	1.00	0.89	1.00

and non-membership value for triangular, trapezoidal and Gaussian functions. The resultant values of these functions are applied to Euclidean n – norm Eq. 15 for performing the data perturbation.

Let's consider age attribute from adult income dataset for data perturbation. The membership function value $\mu_A(x)$ using the Eqs. 4, 5 and 6 has been found and been applied on Euclidean n – norm for data perturbation is shown in the Table 4.

Table 5 shows the data perturbation value using intuitionistic fuzzy set. Here ITMF and ITrMF are used to yield the membership and non-membership values along with the hesitation index. These resultant values are given as the input for Euclidean n – norm formula (Eq. 15) to perform the data perturbation. The Gaussian function with IFS has discussed separately as it is main focused area of research.

5.3 Phase 2—Intuitionistic fuzzy Gaussian membership function (IGMF) for data perturbation

Gaussian function for intuitionistic fuzzy has specified with two arguments. It has a central value m and width k , where $k > 0$. The Gaussian curve will be narrower if the value of k is small.

The Gaussian membership function and non-membership function for the intuitionistic fuzzy is represented in the Eqs. 16 and 17.

$$\mu_A(x) = \exp(-(x - m)^2 / (2(k)^2)) \quad (16)$$

$$\nu_A(x) = 1 - \exp(-(x - m)^2 / (2(k)^2)) \quad (17)$$

The hesitation index for the Eqs. (16) and (17) is,

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \quad (18)$$

Here, x denotes the membership value.

In the above equation, the hesitation index $\pi_A(x)$ value becomes zero. If $\pi_A(x) = 0$, then it will be equivalent to a normal fuzzy set. For obtaining the intuitionistic fuzzy, the least value which will earn the highest perturbation level is considered as the value for $\pi_A(x)$. In this research, the value 0.1 yields a good perturbation level than other values when applied for $\pi_A(x)$.

Let's consider the adult income dataset for performing the data perturbation using $\pi_A(x)$. As discussed, the hesitation index value for the adult income dataset is assigned to be 0.1 (i.e.) $\pi_A(x) = 0.1$. Based on the hesitation index $\pi_A(x) = 0.1$, the membership and non-membership function can be re-written as,

$$\mu_A(x) = \exp(-(x - m)^2 / (2(k)^2)) - 0.1 \quad (19)$$

Table 6 Data perturbation for age attribute from adult income dataset after calculating IV by applying IGMF

Operation	Age attribute				
	39	50	38	53	28
Original values					
After calculating IV	7.17	7.28	1.90	2.68	7.11
Membership degree for IGMF (Eq. 19)	0.99	0.70	0.99	0.58	0.74
Non membership degree for IGMF (Eq. 20)	– 0.09	0.20	– 0.09	0.32	0.16
Hesitation index (Eq. 18)	0.1	0.1	0.1	0.1	0.1
Euclidean 2- norm (Eq. 15)	1.00	0.74	1.00	0.67	0.76
Perturbed data usingIGMF	1.00	0.74	1.00	0.67	0.76

Table 7 Data perturbation using IGMF for adult income dataset

Operation	Age	Education	Occupation
Original value	39	9th	Sales
After calculating IV	7.17	24.28	46.68
After applying IGMF	1.00	0.77	0.92

$$\nu_A(x) = 0.9 - (\exp(-(x - m)^2 / (2(k)^2))) \quad (20)$$

After finding the values for $\mu_A(x)$, $\nu_A(x)$ and $\pi_A(x)$, the Euclidean n– norm formula given in the Eq. 15 has been used to perform the second phase of data perturbation.

Consider the age attribute from adult income data set. Using the Gaussian function, the membership degree, non-membership degree and hesitation index has been calculated for age attribute and these resultant values are applied in Eq. 15 to perform the data perturbation as shown in Table 6.

In the same way, the age, education and occupation attribute from the adult income datasets have perturbed using the procedure employed in Table 6. The final perturbed value of age, education and occupation attribute is shown in the Table 7.

The Table 7 shows the perturbed value for age, education and occupation attribute from adult income dataset. In Table 7, both the numerical and categorical attributes are chosen for perturbation. The same approach has been applied for bank marketing and lung cancer dataset to perform the data perturbation. If this perturbed data is given to external parties for analysis, it is impossible for them to extract the sensitive and private data about the individual.

5.4 Statistical transformation with intuitionistic fuzzy algorithm

The pseudo-code for performing the data perturbation using WOE, IV, intuitionistic fuzzy Gaussian membership function and Euclidean n-norm is named as statistical

transformation with intuitionistic fuzzy algorithm. The proposed STIF algorithm can be applied for both numerical and categorical data for perturbation.

Algorithm: Statistical Transformation with Intuitionistic Fuzzy (STIF) algorithm.

Input: Attribute X

Random Value R

Output: Perturbed Attribute X'

STIF contains the following pseudo codes.

Pseudo code for WOE & IV:

1. Fragment the attributes into P bins
2. Compute the values for bin values based on the class label
3. Calculate the percentage of instance and non_instance
4. Compute WOE

$$WOE = \ln(\text{instance \%} / \text{non_instance \%})$$
5. Compute IV

$$IV = \sum (\text{instance \%} - \text{non_instance \%}) \times WOE$$
6. Arithmetic operation between IV and random value R

Pseudo code for intuitionistic fuzzy Gaussian membership function:

1. Input IV
2. Intuitionistic fuzzy set A in universal set X
3. Compute the membership function $\mu_A(x)$

$$\mu_A(x) = \exp\left(-\frac{(x-m)^2}{2(k)^2}\right)$$
4. Compute the non membership function $\nu_A(x)$

$$\nu_A(x) = 0.9 - \left(\exp\left(-\frac{(x-m)^2}{2(k)^2}\right)\right)$$
5. Calculate the hesitation index $\pi_A(x)$

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x), \text{ here } \pi_A(x) = 0.1$$

Pseudo code for Euclidean n –norm:

1. Input $\mu_A(x)$, $\nu_A(x)$ and $\pi_A(x)$
2. Calculate Euclidean n -norm, $n=2$

$$|S|_2 = \sqrt{(s_1^2 + s_2^2 \dots + s_n^2)}, S = (S_1, S_2, \dots, S_n)$$
3. $|S| \rightarrow$ Perturbed value

The experimentation using fuzzy set and intuitionistic fuzzy set with triangular, trapezoidal and Gaussian membership functions has been done on the adult income, bank marketing and lung cancer datasets to perform the data perturbation. Among the membership functions, Gaussian membership function with

intuitionistic fuzzy set yields the better result. So, the STIF algorithm is proposed based on IGMF for phase 2 of data perturbation.

6 Results and discussion

This section presents and highlights the implementation results of the proposed STIF algorithm. The STIF algorithm is employed on three benchmark datasets including adult income, bank marketing and lung cancer taken from the famous UCI machine learning repository. The classification models Decision Tree (DT), Extreme Gradient Boost (XGB), Random Forest (RF) and Support Vector Machine (SVM) is used to find the performance and accuracy level of the STIF algorithm. In addition, STIF algorithm has tested for Privacy Preserving Capacity (PPC), Data Perturbation Capacity (DPC) and Data Retrieval Capacity (DRC). The proposed algorithm has compared with state-of-art algorithms like TMF, TrMF, GMF, ITMF, ITrMF, Homomorphic Encryption using ElGamal Algorithm (HEA) and Differential Privacy Scheme using Gaussian Mechanism (DPS). The implementation results show that the STIF algorithm outperforms in accuracy, performance and privacy preserving measures along with 100% of data utility.

Many state-of-art algorithms and methods have been applied only to the numerical attributes. But the proposed algorithm has been applied to both numerical and categorical attributes to perform the data perturbation. It is one of the important highlight of the proposed work.

6.1 Datasets description

Three datasets has been chosen for analysis from UCI machine learning repository. The adult income dataset holds 14 records with 48,842 instances. It contains both the continuous and discrete values with the class label “income”. The income field is divided into two classes: ≤ 50 K and > 50 K. Among 14 records, 6 are continuous attributes and 8 are nominal attributes. Table 8 presents the lists of attributes and their value from the adult income dataset.

Table 8 Attributes and value from adult income dataset

Attributes	Value	Attributes	Value
Age	Continuous	Relationship	Nominal
Workclass	Nominal	Race	Nominal
Fnlwgt	Continuous	Sex	Nominal
Education	Nominal	Capital-gain	Continuous
Education-num	Continuous	Capital-loss	Continuous
Marital-status	Nominal	Hours-per-week	Continuous
Occupation	Nominal	Native-country	Nominal

Table 9 Attributes and value from bank marketing dataset

Attributes	Value	Attributes	Value
Age	Continuous	Duration	Continuous
Job	Nominal	Campaign	Continuous
Marital	Nominal	Pdays	Continuous
Education	Nominal	Previous	Continuous
Default	Nominal	Poutcome	Nominal
Housing	Nominal	Employment variation rate	Continuous
Loan	Nominal	Consumer price index	Continuous
Contact type	Nominal	Consumer confidence index	Continuous
Month	Nominal	Euribor 3 month rate	Continuous
Day_of_week	Nominal	Number of employees	Continuous

The bank marketing dataset has 20 attributes with 45,211 numbers of instances. The classification goal of this dataset is to predict whether the client will subscribe to a term deposit or not (yes/no). Among the 20 attributes, 10 of them have numerical values and 10 attributes have categorical values. Table 9 represents the lists of attributes and their value from the bank marketing dataset.

The lung cancer dataset contains 16 records with 284 instances. The dataset contains three types of pathological lung cancers. All the predictive attributes in the dataset are nominal taking the values from 0 to 3. Among the 16 attributes, 2 are nominal values and 14 attributes possess numerical values. Table 10 represents the lists of attributes and their value from the lung cancer dataset.

All these datasets contains the final class label, so these datasets are categorized under the supervised learning models.

6.2 Accuracy

Accuracy is the measure of closeness between the obtained values to the true values. In general the classification accuracy can be measured using the Eq. 21.

Table 10 Attributes and value from lung cancer dataset

Attributes	Value	Attributes	Value
Gender	Nominal	Allergy	Continuous
Age	Continuous	Wheezing	Continuous
Area	Nominal	Alkohol	Continuous
Smoking	Continuous	Fatigue	Continuous
Yellow fingers	Continuous	Coughing	Continuous
Anxiety	Continuous	Shortness of breath	Continuous
Peer_pressure	Continuous	Swallowing difficulty	Continuous
Chronic disease	Continuous	Chest pain	Continuous

Table 11 Accuracy for adult income dataset using FS and IFS

Classifier models	Accuracy in %								
	Original dataset	TMF	TrMF	GMF	ITMF	ITrMF	HEA	DPS	STIF
DT	72.25	70.8	72.25	72.25	72.25	72.25	71.31	71.44	72.25
XGB	78.25	76.25	76.1	76.75	76.35	76.4	73.78	74.54	76.9
RF	76.35	75.05	75.35	75.7	75.45	75.14	72.60	73.01	76
SVM	75.05	74.14	74.2	74.23	74.35	73.95	71.32	71.98	74.8

Table 12 Accuracy for bank marketing dataset using FS and IFS

Classifier models	Accuracy in %								
	Original dataset	TMF	TrMF	GMF	ITMF	ITrMF	HEA	DPS	STIF
DT	71	71	71	71	71	71	70.43	70.86	71
XGB	74.25	73.45	73.35	73.75	73.9	73.9	71.89	72.06	74.2
RF	74.85	74.24	74.25	74.85	74.37	74.35	72.34	72.77	74.85
SVM	69.2	68.12	68.7	69	68.75	68.25	66.63	67.10	69.1

Table 13 Accuracy for lung cancer dataset using FS and IFS

Classifier models	Accuracy in %								
	Original dataset	TMF	TrMF	GMF	ITMF	ITrMF	HEA	DPS	STIF
DT	87.5	87.5	87.5	87.5	87.5	87.5	81.37	82.05	87.5
XGB	100	87.5	87.5	100	87.5	87.5	85.11	85.66	100
RF	100	100	100	100	100	100	88.96	89.23	100
SVM	100	85.7	85.7	92.8	85.7	92.8	82.65	84.25	100

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (21)$$

where TP – True Positive, TN – True Negative, FP – False Positive and FN – False Negative.

The accuracy for adult income, bank marketing and lung cancer dataset has been calculated based on the Eq. 18 and it is shown in the Tables 11, 12, and 13 respectively. The Tables 11, 12 and 13 compares the state-of-art algorithms and the proposed STIF algorithm based on the accuracy. The classification models DT, XGB, RF and SVM are used for accuracy analysis on the respective datasets.

Tables 11, 12 and 13 shows the percentage of accuracy obtained by the classifiers DT, XGB, RF and SVM on adult income, bank marketing and lung cancer datasets. From the table, the proposed STIF algorithm outperforms all other

approaches. The accuracy of STIF is nearly equal to the original dataset. The overall accuracy of the STIF algorithm is (100, 98.3, 99.5, 99.7), (100, 99.9, 100, 99.9) and (100,100,100,100) for adult income, bank marketing and lung cancer datasets respectively using the classifier models. The STIF algorithm has the higher accuracy percentage than all the fuzzy set approaches and intuitionistic fuzzy sets with trapezoidal and triangular functions.

6.3 Data perturbation capacity (DPC)

The STIF algorithm has applied on adult income, bank marketing and lung cancer datasets for data perturbation. It is more important to measure the DPC of the STIF algorithm.

The DPC has been measured using the variance. The variance is the measure of variability. It measures how faraway the set of numbers extended out from their mean value. The formula for the variance is given in the Eq. 22.

$$\text{Var}(X) = \frac{1}{p} \sum_{i=1}^p (x_i - \mu)^2 \quad (22)$$

where X denotes the random value, $x_n \rightarrow p_n$ represents the probability mass function and μ represents the mean value and it is given by

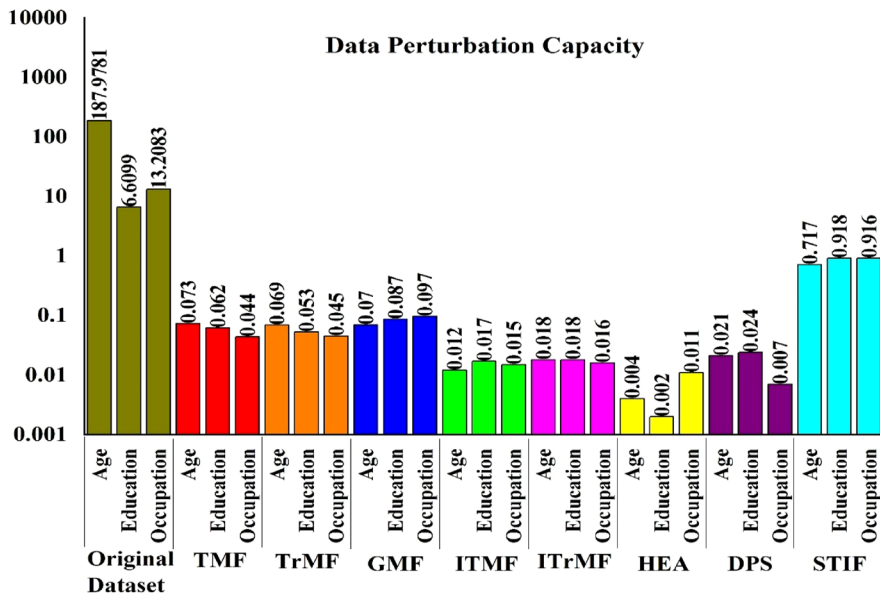


Fig. 2 DPC for adult income dataset

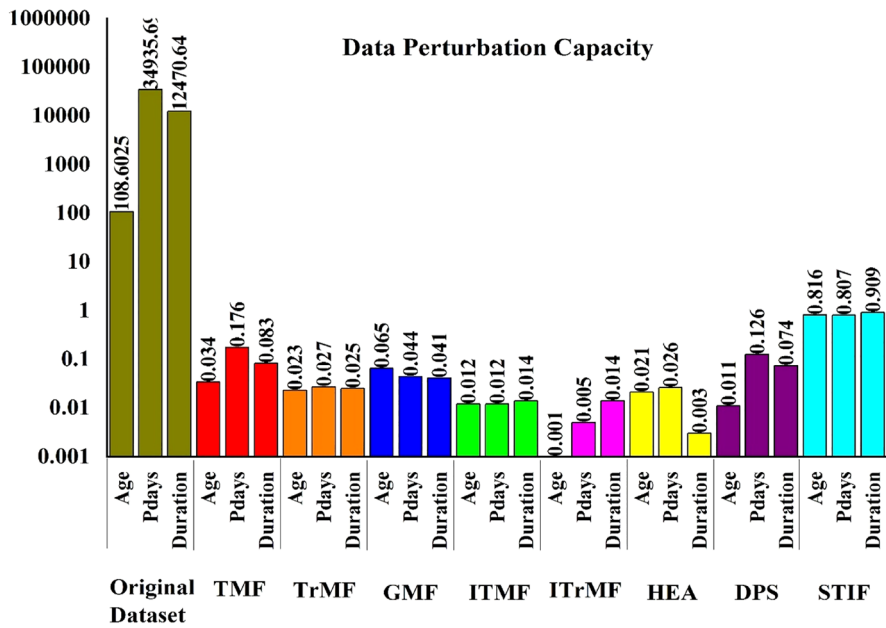


Fig. 3 DPC for bank marketing dataset

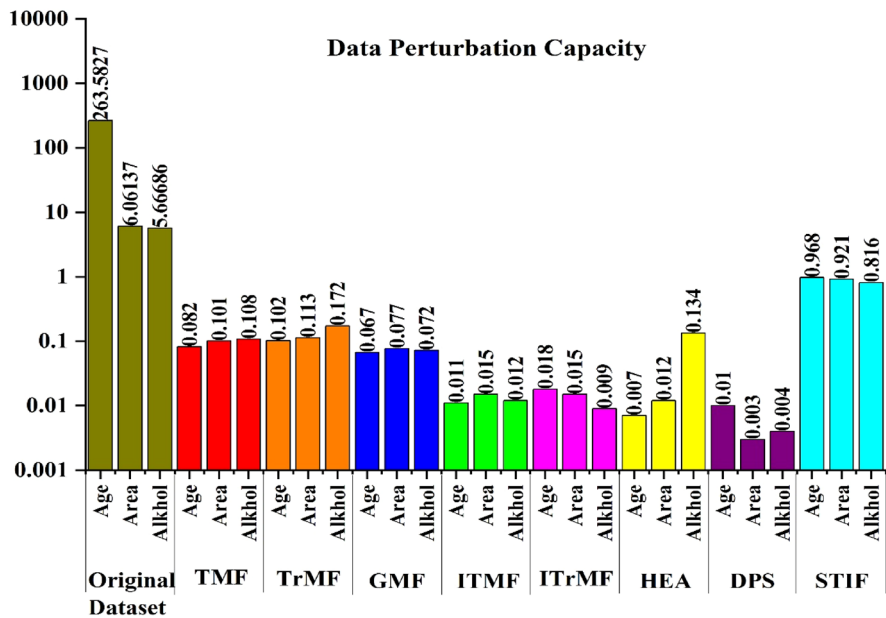


Fig. 4 DPC for lung cancer dataset

$$\mu = \frac{1}{p} \sum_{i=1}^p x_i. \quad (23)$$

The Eq. 22 is applied for all the original dataset and the STIF applied perturbed dataset to find the variance. Higher variance value results in higher data perturbation. Figure 2, 3 and 4 shows the DPC of adult income, bank marketing and lung cancer datasets respectively.

The attributes (age, education, occupation) from adult income dataset, (age, pdays, duration) from bank marketing dataset and (age, area, alkhoh) from lung cancer dataset are chosen for calculating the data perturbation capacity. The TMF, TrMF, GMF, ITMF, ITrMF, HEA, DPS and STIF algorithms were applied on those attributes in the datasets for perturbation. Higher variance results in higher data perturbation. From the Figs. 2, 3 and 4 it is more evident that the variance of STIF algorithm is higher than other algorithms. The variance of STIF algorithm is (0.717, 0.918 and 0.916) for age, education and occupation attributes for adult income dataset, (0.816, 0.807 and 0.909) for age, pdays, duration attributes for bank marketing dataset and (0.968, 0.921 and 0.816) for age, area and alkhoh attributes for lung cancer dataset respectively. So, STIF has the highest capability to perturb the data. In addition, the variance of all the three original dataset is much higher than the state-of-art approaches and the STIF algorithm, because the fuzzy set and intuitionistic fuzzy set values will be in the range of 0 to 1.

6.4 Privacy preserving capacity (PPC)

Another important measure in the data privacy is the PPC. The algorithms should preserve the privacy of an individual's to the greater extent along with data perturbation level. The PPC can be measured using the extended version of variance metric as shown in the Eq. 24.

$$\text{Privacy Preserving Capacity} = (\text{Var}(R - M)) / (\text{Var}(R)) \quad (24)$$

where R — real / actual dataset, M — modified / perturbed dataset.

The PPC for TMF, TrMF, GMF, ITMF, ITrMF, HEA, DPS and STIF are calculated based on the Eq. 24 by employing the algorithms in the attributes from adult income, bank marketing and lung cancer datasets. The higher value for Eq. 24 results in higher privacy preserving capacity.

The Figs. 5, 6 and 7 show the PPC of adult income, bank marketing and lung cancer datasets respectively. From the figures, the PPC value of STIF algorithm is higher than other algorithms. The PPC of STIF algorithm is (0.997, 1.039 and 1.004) for age, education and occupation attributes for adult income dataset, (0.999, 1.012 and 0.97) for age, pdays and duration attributes for bank marketing dataset and (1.007, 0.989 and 0.997) for age, area and alkhoh attributes for lung cancer dataset respectively. By comparing the STIF algorithm, ITrMF and TMF stands next in the PPC for adult income dataset, TrMF and ITMF gives the second and third level

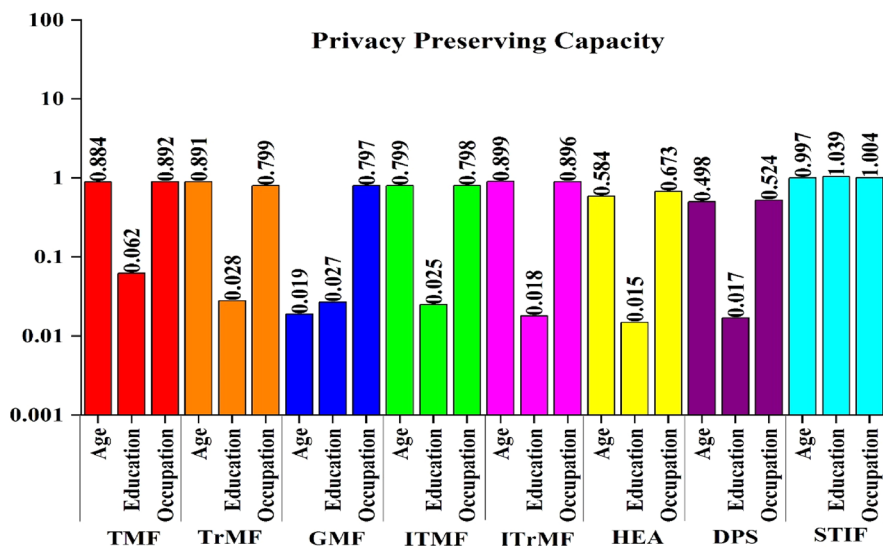


Fig. 5 PPC for adult income dataset

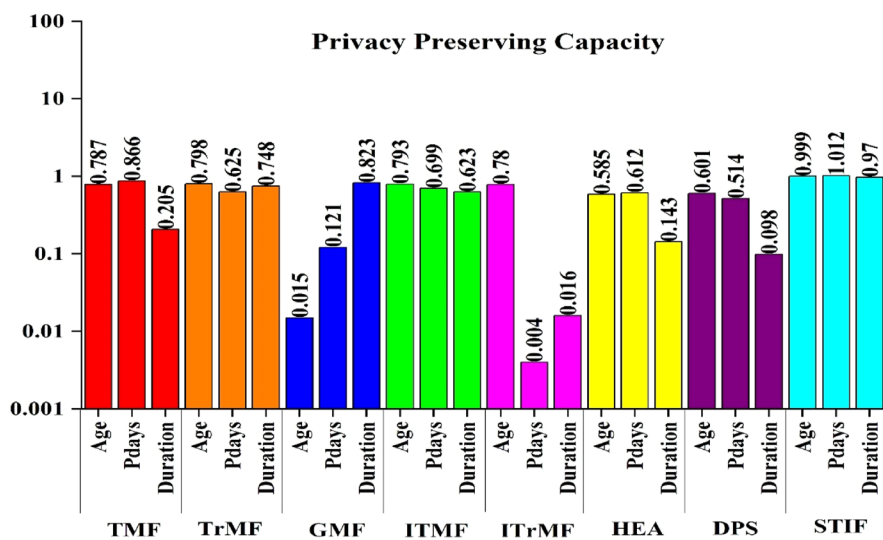


Fig. 6 PPC for bank marketing dataset

of PPC for bank marketing dataset and TMF and ITMF provides the next level of PPC for lung cancer dataset. It is more evident that the proposed STIF algorithm provides the highest PPC values for all the three datasets. This indicates that the STIF outperforms in privacy preservation than the state-of-art algorithms.

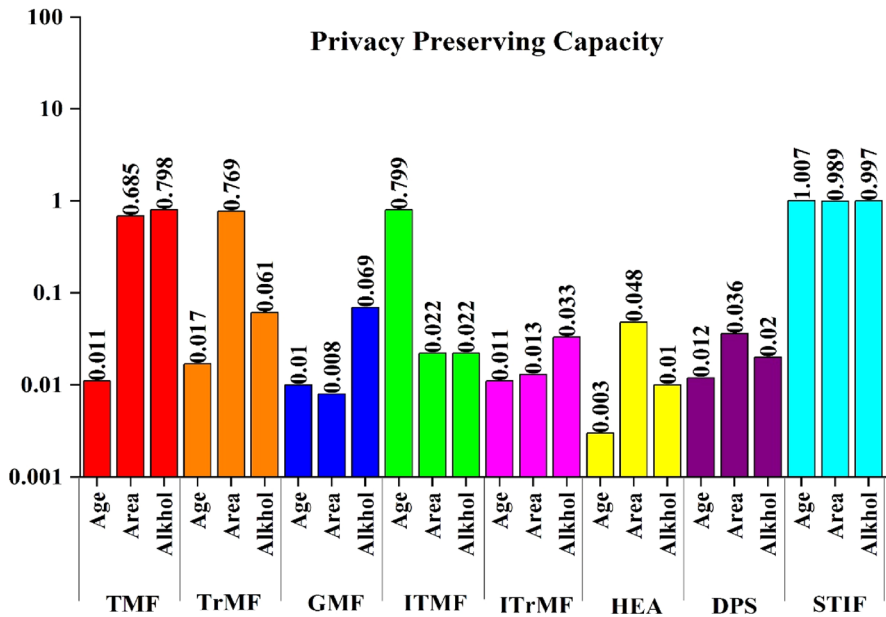


Fig. 7 PPC for lung cancer dataset

6.5 Data retrieval capacity (DRC)

The data perturbation results in data alteration. After perturbation, it should also be possible to recover or retrieve the original data again on need basis. There should not be any information loss. The algorithm should preserve the data privacy as well as the data utility.

DRC plays a major role in preserving the data utility and retrieve back the original data. The metrics precision and recall are used to find the DRC of an algorithm. Precision and recall are based on the data relevance. Precision is the division of related events among the recovered events, while the recall is the division of related events that are recovered.

$$\text{Precision} = \frac{|\{\text{related events}\} \cap \{\text{recovered events}\}|}{|\{\text{recovered events}\}|} = P(\text{related}|\text{recovered}) \quad (25)$$

Table 14 Contingency table for precision and recall

	Related	Non-related
Related	TP	FP
Non-related	FN	TN

$$\text{Recall} = \frac{|\{\text{related events}\} \cap \{\text{recovered events}\}|}{|\{\text{related events}\}|} = P(\text{recovered}|\text{related}) \quad (26)$$

The Eqs. 25 and 26 can be made clear by inspecting the succeeding contingency table.

Based on Table 14, the Eqs. 25 and 26 can be re-written as,

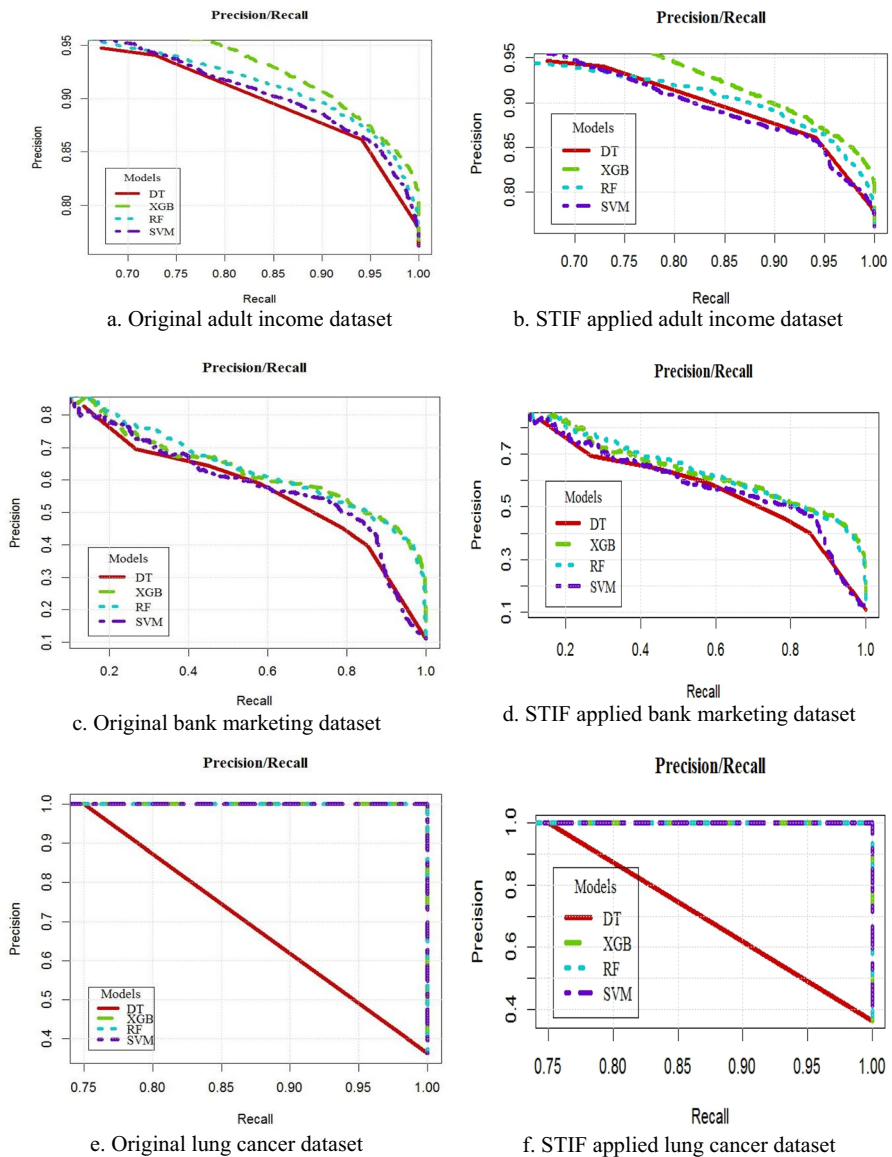


Fig. 8 Precision/Recall plot for actual and perturbed dataset using STIF algorithm

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (27)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (28)$$

The resultant values from TMF, TrMF, GMF, ITMF, ITrMF, HEA, DPS and STIF employed datasets are applied on the Eqs. 27 and 28 to find the precision and recall values. The classifier DT, XGB, RF and SVM are used for the analysis to predict the related and recovered events.

Figure 8 shows the precision and recall plot for original dataset and STIF applied perturbed dataset. The precision and recall plots are nearly same for both datasets. This indicates that, the knowledge extracted before data perturbation and after data perturbation is nearly same. STIF has not produced any information loss on all the datasets and thus preserves the data utility. The precision and recall is above 95% in adult income dataset, above 85% in bank marketing dataset and 100% in lung cancer dataset in accordance with the original dataset.

In addition, the classifier DT, XGB, RF and SVM gives the same order of plots in all the datasets. From the lowest to highest order in the proposed classification are DT, SVM, RF and XGB. DT gives the less classification result, while XGB yields the highest classification result. This hierarchical order is replicated in all the three STIF applied datasets. It proves that the utility of the data is preserved along with data privacy. The original data can be retrieved even after the data perturbation on a needed basis. This will not affect the mining result.

6.6 Performance measure

Beyond the accuracy, the performance and validity of an algorithm should be measured by the statistical measure such as sensitivity and specificity. Sensitivity is the true positive rate which measures the percentage of positives that are exactly recognised. In turn, Specificity is the true negative rate which measures the percentage of negatives that are recognised exactly by an algorithm.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (29)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (30)$$

The terms true positives, true negatives, false positives and false negatives states the results of an algorithm and the correctness of the classification techniques. The proposed STIF algorithm has been applied on classifiers DT, XGB, RF and SVM to analyse the performance using adult income, bank marketing and lung cancer datasets. The resultant values from the each dataset using STIF algorithm is applied in Eqs. 29 and 30 to get the sensitivity–specificity curves as shown in the Fig. 9.

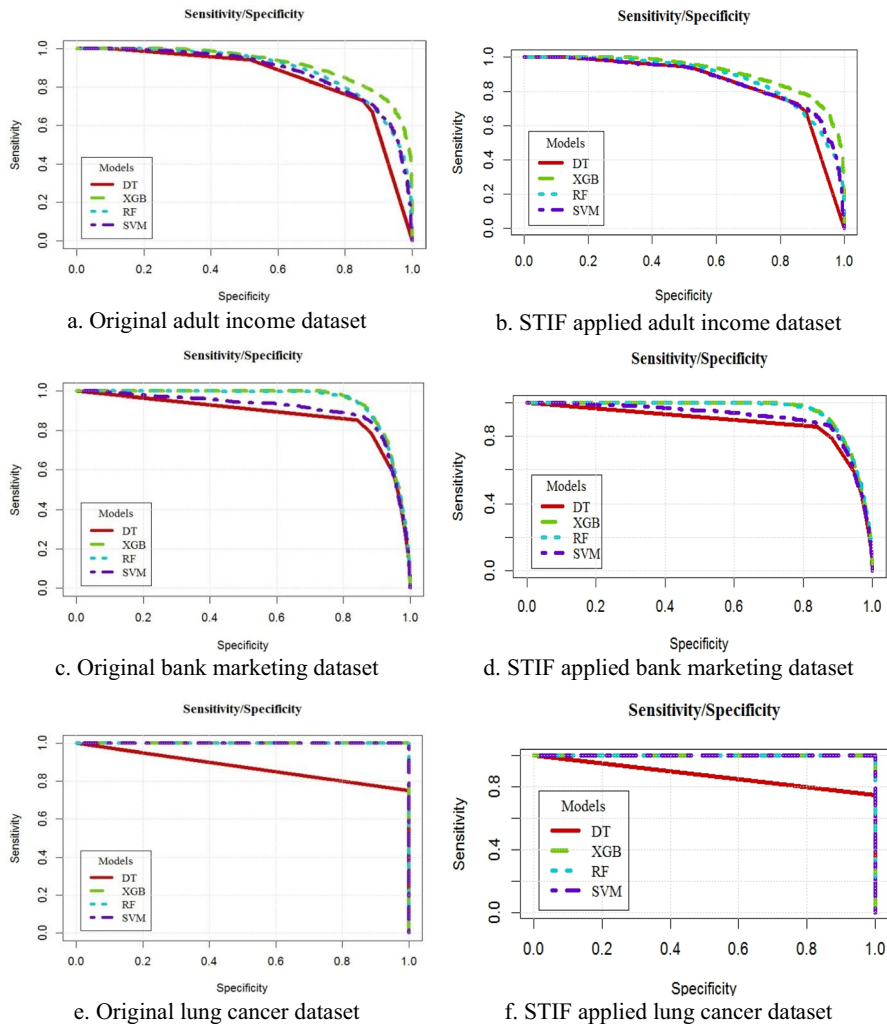


Fig. 9 Sensitivity/Specificity plot for actual and perturbed dataset using STIF algorithm

The Fig. 9a, c, and e portrays the sensitivity–specificity curve of original datasets and Fig. 9b, d and f displays the sensitivity–specificity curve of STIF applied adult income, bank marketing and lung cancer datasets respectively. The plots in the Fig. 9 show that the sensitivity and specificity curve for both original and perturbed datasets remains equal which indicates the higher performance of STIF algorithm. The sensitivity and specificity values are 100% for adult income dataset and above 90% for bank marketing and lung cancer datasets respectively.

Further, the pattern extracted from the original dataset is remains same in the STIF applied perturbed dataset. This again shows that the STIF algorithm preserves the private and sensitive data of an individual along with the data utility.

This perturbed data can be given for any external parties for analysis. They cannot extract any private and sensitive information about the individual which is the key success of the proposed work.

6.6.1 Performance measure at various thresholds

Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) is used to measure the performance of the classification problems at the various threshold level. ROC is the probability curve and AUC symbolizes the amount of separability. AUC is the summary of ROC curve. The higher value of AUC results in greater performance at different threshold levels.

The ROC and AUC curve can be plotted with true positive rate (TPR) in y-axis against the false positive rate (FPR) in x-axis. In the proposed work, the ROC-AUC is plotted for the classifiers DT, XGB, RF and SVM in adult income, bank marketing and lung cancer datasets between the values 0 to 1. The formula to plot the ROC and AUC is given in the Eqs. 31 and 32.

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \quad (31)$$

Table 15 AUC for Adult income dataset

Classifier	AUC for actual dataset	AUC for perturbed dataset using STIF algorithm
DT	0.84	0.84
XGB	0.92	0.91
RF	0.89	0.88
SVM	0.88	0.87

Table 16 AUC for bank marketing dataset

Classifier	AUC for actual dataset	AUC for perturbed dataset using STIF algorithm
DT	0.88	0.88
XGB	0.95	0.95
RF	0.95	0.95
SVM	0.91	0.91

Table 17 AUC for lung cancer dataset

Classifier	AUC for actual dataset	AUC for perturbed dataset using STIF algorithm
DT	0.88	0.88
XGB	1	1
RF	1	1
SVM	1	1

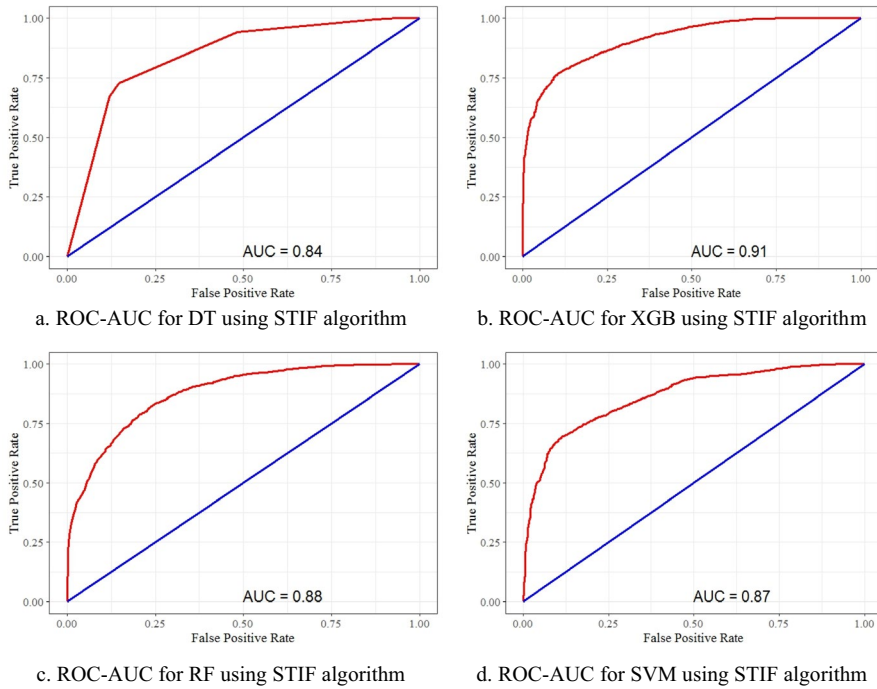


Fig. 10 ROC and AUC for the perturbed adult income dataset using STIF algorithm

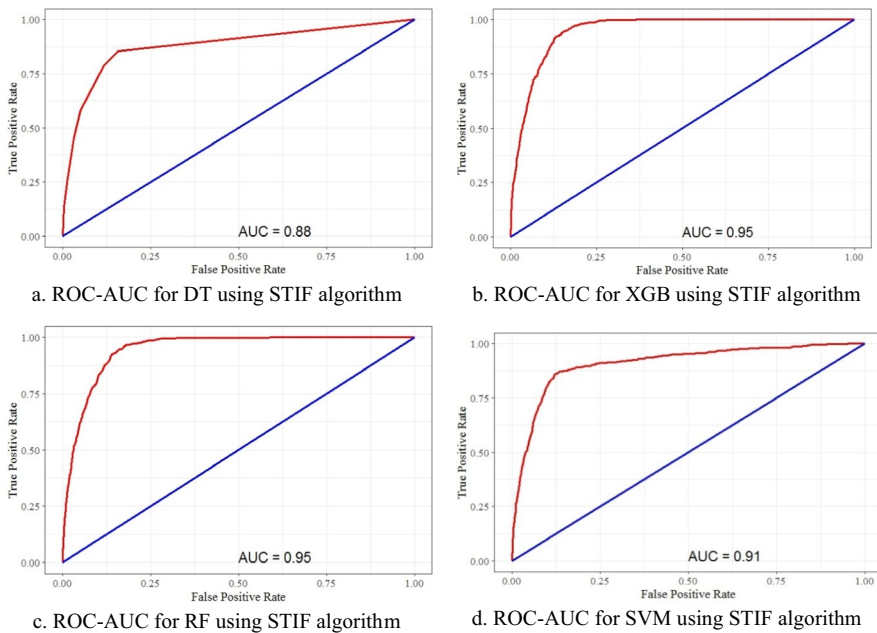


Fig. 11 ROC and AUC for the perturbed bank marketing dataset using STIF algorithm

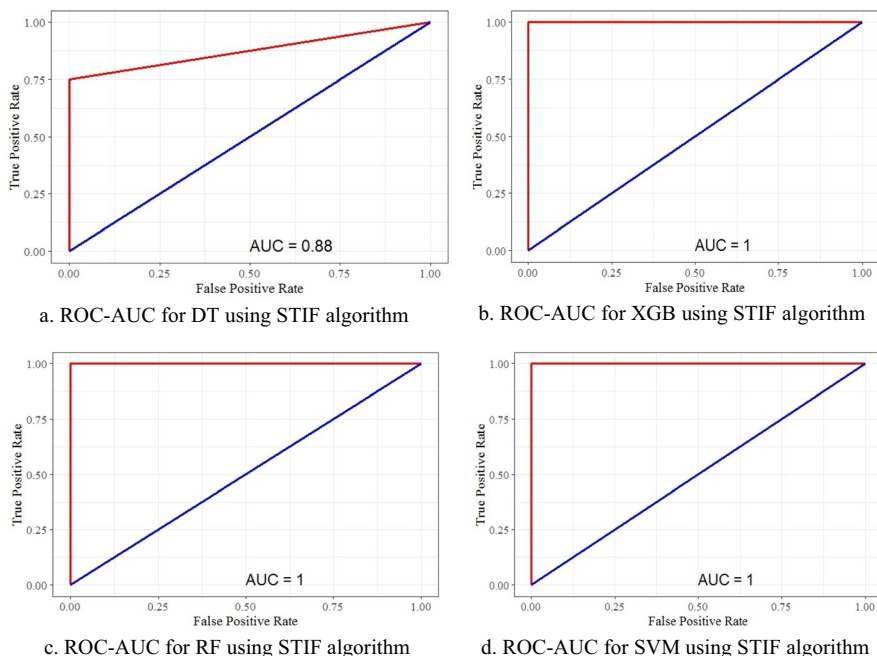


Fig. 12 ROC and AUC for the perturbed lung cancer dataset using STIF algorithm

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) \quad (32)$$

Tables 15, 16 and 17 shows the AUC value for actual and STIF applied perturbed datasets. The values are nearly equal for both datasets. The higher value for AUC results in higher performance. Here, all the values for AUC are much higher and equal to the actual dataset. This indicates the better performance of the STIF algorithm at various threshold limits than the state-of-art algorithms.

Figures 10, 11 and 12 shows the ROC-AUC curve of adult income, bank marketing and lung cancer datasets respectively. The AUC value for all the datasets using the classifier models DT, XGB, RF and SVM is (0.84, 0.91, 0.88 and 0.87), (0.88, 0.95, 0.95 and 1) and (0.88, 1, 1 and 1) respectively. The average values of ROC-AUC for the datasets are 87%, 95% and 97% for adult income, bank marketing and lung cancer datasets respectively using the classifier models.

6.7 Overall comparison of STIF algorithm with state-of-art algorithms

Privacy preservation is more important in this technological era. The proposed STIF algorithm has developed using the statistical transformation methods weight of evidence, information value and intuitionistic fuzzy Gaussian membership function. The accuracy, data perturbation level and privacy preserving level of the STIF algorithm

Table 18 Accuracy comparison of STIF algorithm with the state-of-art algorithms

	Adult Income Dataset				Bank marketing dataset				Lung cancer dataset			
	Accuracy (%)				Accuracy (%)				Accuracy (%)			
	DT	XGB	RF	SVM	DT	XGB	RF	SVM	DT	XGB	RF	SVM
TMF	70.8	76.2	75.0	74.1	71	73.4	74.2	68.1	87.5	87.5	100	85.7
TrMF	72.2	76.1	75.3	74.2	71	73.3	74.2	68.7	87.5	87.5	100	85.7
GMF	72.2	76.7	75.7	74.2	71	73.7	74.8	69	87.5	100	100	92.8
ITMF	72.2	76.3	75.4	74.3	71	73.9	74.3	68.7	87.5	87.5	100	85.7
ITrMF	72.2	76.4	75.1	73.9	71	73.9	74.3	68.2	87.5	87.5	100	92.8
HEA	71.3	73.7	72.6	71.3	70.4	71.8	72.3	66.6	81.3	85.1	88.9	82.6
DPS	71.4	74.5	73.0	71.9	70.8	72.0	72.7	67.1	82.0	85.6	89.2	84.2
STIF	72.25	76.9	76	74.8	71	74.2	74.85	69.1	87.5	100	100	100

outperform the state-of-art algorithms. Table 18 shows the accuracy comparison of STIF algorithm and the state-of-art algorithms.

From the Table 18, it is clear that the accuracy percentage of the proposed STIF algorithm is higher than the state-of-art algorithms. Now, the STIF applied perturbed data can be given for the external parties for analysis. The external parties cannot extract the private and sensitive data about the individual. Thus the STIF algorithm preserves the privacy of an individual by giving higher accuracy, data perturbation capacity and privacy preserving capacity. Further STIF perturbed the data without any data loss while mining the original data and thus preserves the data utility.

7 Conclusion

Data mining provides better service and quality to the today's business needs at lower cost. It exposes the private and sensitive data about the individual to the external parties. It creates a serious threat to the personal data. In this paper, the issue of PPDM has been addressed using the proposed STIF algorithm. The STIF algorithm contains the statistical transformation method weight of evidence, information value and intuitionistic fuzzy Gaussian membership function for data perturbation. The STIF algorithm has applied on adult income, bank marketing and lung cancer datasets. The classifier models DT, XGB, RF and SVM are used for analysing the STIF algorithm. The experimental results show the accuracy of STIF algorithm is 99% in adult income dataset and 100% in both bank marketing and lung cancer datasets. The data perturbation capacity and privacy preserving capacity of the STIF algorithm has measured by the variance metric and STIF outperforms in both DPC and PPC comparing with the state-of-art algorithms. The data retrieval capacity of STIF algorithm is 95%, 85% and 100% for adult income, bank marketing and lung cancer datasets respectively. Sensitivity–Specificity and ROC-AUC has been taken as the performance measure of STIF. The obtained sensitivity–specificity plot for both the original and STIF applied perturbed dataset is nearly equal for all the datasets. The

ROC curve is plotted between 0 to 1, the STIF obtains the AUC values as 0.91, 0.95 and 1 for adult income, bank marketing and lung cancer datasets respectively. Higher AUC value results in higher algorithmic performance. The main advantage of the proposed STIF algorithm is that it can be applied to both numerical and categorical data. Also, the STIF algorithm can be employed for different types of datasets like bivariate datasets, correlation datasets and transactional datasets. In addition, the pattern extracted before data perturbation and after STIF applied perturbation is identical. This shows that there is no loss of information after data perturbation and thus preserves the data utility. The results indicate that the STIF algorithm preserves the private and sensitive data of an individual to the greater extent along with data utility. The proposed work can be applied to real-world areas like healthcare, finance, sales, education and other business. These domains contain many private and sensitive data about the customers. So, it is the duty of the service providers to assure the data privacy of their own customers. The STIF algorithm is like a ray of sunshine for the service providers to provide a guarantee to preserve the client's data privacy.

Author contributions G. Sathish Kumar: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Writing - original draft, Writing - review & editing, Carrying out additional analyses. K. Premalatha: Formal analysis, Methodology Investigation, Supervision, Validating the ideas, Carrying out additional analyses, Reviewing this paper.

Funding Not applicable.

Data availability Datasets used in this work has been taken from the publically available repository (UCI machine learning repository).

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest No conflict of interests.

Ethical approval Not Applicable.

References

1. Yongjian, F.: Data mining. *IEEE Potentials* **16**(4), 18–20 (1997). <https://doi.org/10.1109/45.624335>
2. Aggarwal, C.C., Yu, P.S.: Privacy-preserving data mining: models and algorithms. *Adv. Database Syst.* **34**(22), 514 (2008)
3. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In: *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York, USA, pp. 247–255 (2001)
4. Liu, J., Huang, B., Li, H., Bu, X., Zhou, X.: Optimization-based three-way decisions with interval-valued intuitionistic fuzzy information. *IEEE Trans. Cybern.* (2022). <https://doi.org/10.1109/TCYB.2022.3151899>

5. Zhang, Q., Cheng, Y., Zhao, F., Wang, G., Xia, S.: Optimal scale combination selection integrating three-way decision with Hasse diagram. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(8), 3675–3689 (2022). <https://doi.org/10.1109/TNNLS.2021.3054063>
6. Ye, X., Liu, D.: A cost-sensitive temporal-spatial three-way recommendation with multi-granularity decision. *Inf. Sci.* **589**, 670–689 (2022). <https://doi.org/10.1016/j.ins.2021.12.105>
7. Atanassov, K.T.: Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **20**(1), 87–96 (1986)
8. Atanassov, K.T.: Type-1 fuzzy sets and intuitionistic fuzzy sets. *Algorithms* **10**(3), 1–12 (2017)
9. Qian, G., Wang, H., Feng, X.: Generalized hesitant fuzzy sets and their application in decision support system. *Knowl.-Based Syst.* **37**(1), 357–365 (2013)
10. Soloshenko, O.M.: Generalizations of logistic regression, weight of evidence, and the gini index for a continuous target variable taking on probabilistic values. *Cybern. Syst. Anal.* **51**(6), 992–1004 (2015)
11. Agrawal, R., Srikant, R., Thomas, D.: Privacy-preserving OLAP. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of data*, pp. 251–262 (2005). <https://doi.org/10.1145/1066157.1066187>
12. Sarala, V., Shanmugapriya, P.: DLFPM-SSO-PE: privacy-preserving and security of intermediate data in cloud storage. *Distrib. Parallel Databases* **40**(4), 815–833 (2022). <https://doi.org/10.1007/s10619-021-07352-z>
13. Yi, X., Zhang, Y.: Privacy-preserving distributed association rule mining via semi-trusted mixer. *Data Knowl. Eng.* **63**(2), 550–567 (2002)
14. Li, Y., Jiang, Z.L., Yao, L., Wang, X., Yiu, S., Huang, Z.: Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties. *Clust. Comput.* **22**(1), 1581–1593 (2019)
15. Zehtabchi, S., Daneshpour, N., Safkhani, M.: A new method for privacy preserving association rule mining using homomorphic encryption with a secure communication protocol. *Wirel. Netw.* (2022). <https://doi.org/10.1007/s11276-022-03185-5>
16. Yao, L., Chen, Z., Hu, H., et al.: Sensitive attribute privacy preservation of trajectory data publishing based on l-diversity. *Distrib. Parallel Databases* **39**(3), 785–811 (2021). <https://doi.org/10.1007/s10619-020-07318-7>
17. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **10**(5), 557–570 (2002)
18. Tan, Y., Wu, W., Liu, J., Wang, H., Xian, M.: Lightweight edge-based KNN privacy-preserving classification scheme in cloud computing circumstance. *Concurr. Comput.* **32**(19), e5804 (2020)
19. Honda, K., Kawano, A., Notsu, A., Ichihashi, H.: A fuzzy variant of k-member clustering for collaborative filtering with data anonymization. In: *Proceedings of the 2012 IEEE International Conference on Fuzzy Systems*, pp. 1–6 (2012). <https://doi.org/10.1109/FUZZ-IEEE.2012.6250782>
20. Kumar, P., Varma, K.I., Sureka, A.: Fuzzy based clustering algorithm for privacy preserving data mining. *Int. J. Bus. Inf. Syst.* **7**(1), 27–40 (2011)
21. Geetha, M.A.: Fuzzy-based random perturbation for real world medical datasets. *Int. J. Telemed. Clin. Pract.* **1**(2), 111–124 (2015)
22. Kumar, S., Kumar, P.: Privacy preserving in online social networks using fuzzy rewiring. *IEEE Trans. Eng. Manag.* **70**, 1–9 (2021). <https://doi.org/10.1109/TEM.2021.3072812>
23. Krishnamoorthy, S., SudhaSadasivam, G., Rajalakshmi, M., et al.: Privacy preserving fuzzy association rule mining in data clusters using particle swarm optimization. *Int. J. Intell. Inf. Technol.* **13**(2), 1–20 (2017)
24. Gupta, M., Joshi, R.C.: Privacy preserving fuzzy association rules hiding in quantitative data. *Int. J. Comput. Theory Eng.* **1**(4), 382–388 (2009)
25. Zdravevski, E., Lameski, P., Kulakov, A.: Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms. In: *Proceedings of the 2011 International Joint Conference on Neural Networks*, vol. 1, no.1, pp. 181–188 (2011). <https://doi.org/10.1109/IJCNN.2011.6033219>
26. Geetha, M.A., Acharjya, D.P., Iyengar, N.C.: Privacy preservation in fuzzy association rules using rough set on intuitionistic fuzzy approximation spaces and DSR. *Int. J. Auton. Adapt. Commun. Syst.* **10**(1), 67–87 (2017)
27. Xia, M.M., Xu, Z.S.: Hesitant fuzzy information aggregation in decision making. *Int. J. Approx. Reason.* **52**(3), 395–407 (2011)
28. Torra, V., Miyamoto, S., Endo, Y., Domingo-Ferrer, J.: On intuitionistic fuzzy clustering for its application to privacy. In: *Proceedings of the 2008 IEEE International Conference on Fuzzy*

- Systems (IEEE World Congress on Computational Intelligence), pp 1042–1048 (2008). <https://doi.org/10.1109/FUZZY.2008.4630498>
29. Banerjee, S., Roy, T.K.: Solution of single and multi objective stochastic inventory models with fuzzy cost components by intuitionistic fuzzy optimization technique. *Adv. Oper. Res.* **2010**(1), 1–19 (2010)
 30. Saranya, K., Premalatha, K.: Privacy-preserving data publishing based on sanitized probability matrix using transactional graph for improving the security in medical environment. *J. Supercomput.* **76**, 5971–5980 (2020). <https://doi.org/10.1007/s11227-019-03102-2>
 31. Radhika, C., Parvathi, R.: Intuitionistic fuzzification functions. *Glob. J. Pure Appl. Math.* **12**(2), 1211–1227 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.