Improved 3D Rotation-based Geometric Data Perturbation Based on Medical Data Preservation in Big Data

Jayanti Dansana[1]*, Dr. Manas Ranjan Kabat[2], Dr. Prasant Kumar Pattnaik[3]
Professor[2, 3]
School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha-751024, India[1, 3]
Department of Computer Science and Engineering, VSSUT Burla, Sambalpur, Odisha-768018, India[2]

***Abstract*—With the rise in technology, a huge volume of data is being processed using data mining, especially in the healthcare sector. Usually, medical data consist of a lot of personal data, and third parties utilize it for the data mining process. Perturbation in health care data highly aids in preventing intruders from utilizing the patient's privacy. One of the challenges in data perturbation is managing data utility and privacy protection. Medical data mining has certain special properties compared with other data mining fields. Hence, in this work, the machine learning (ML) based perturbation approach is introduced to provide more privacy to healthcare data. Here, clustering and IGDP-3DR processes are applied to improve healthcare privacy preservation. Initially, the dataset is pre-processed using data normalization. Then, the dimensionality is reduced by SVD with PCA (singular value decomposition with Principal component analysis). Then, the clustering process is performed by IFCM (Improved Fuzzy C means). The high-dimensional data are divided into several segments by IFCM, and every partition is set as a cluster. Then, improved Geometric Data Perturbation (IGDP) is used to perturb the clustered data. IGDP is a combination of GDP with 3D rotation (3DR). Finally, the perturbed data are classified using a machine learning (ML) classifier - kernel Support Vector Machine- Horse Herd Optimization (KSVM-HHO) to classify the perturbed data and ensure better accuracy. The overall evaluation of the proposed KSVM-HHO is carried out in the Python platform. The performance of the IGDP-KSVM-HHO is compared over the two benchmark datasets, Wisconsin prognostic breast cancer (WBC) and Pima Indians Diabetes (PID) dataset. For the WBC dataset, the proposed method obtains an overall accuracy of 98.08% perturbed data, and for the PID dataset, the proposed method obtains an overall accuracy of 98.04%.**

Keywords*—Data mining; privacy; health care data; machine learning; perturbation; improved fuzzy c-means; horse herd optimization; kernel based support vector machine

I. INTRODUCTION

Big data analysis with proper security and privacy preserving features is required to balance privacy and utility. Recently, the advancement in information science has led to the digitized collection of data and storage in large numbers [1]. HI (health information) system can handle all the information related to health care. The HI system can be operated through national statistics offices and health information departments. It allows data on risk factors related to morbidity, disease, mortality and health service coverage. The quality of patient treatment can be improved by using healthcare software [2]. Some HI elements are indicators, data sources, data management, information products and dissemination. Data mining is a technology that processes a large amount of data to draw unknown, hidden, and potentially useful data from source information [3]. It is used to help a wide range of business applications such as store layout, targeted marketing and customer profiling. Data mining tasks are divided into two categories: predictive tasks and descriptive tasks [4]. Personal and sensitive information of patients may be theft and breaches the patient's privacy through several attacks. A third person may use the information at various levels; hence, the data must be saved. Privacy is essential in sharing information for applications based on knowledge [5].

Recent advances in internet of medical things (IoMT) [6] like ambulance, immediate mobile medical devices faces high delay and low throughput due to the continuous motion state. For overcoming this, resource efficient flow enabled distributed mobility (FDMA) approach is introduced for improving the network performance. In [7], the authors undertaken the case study for latest advancement in health care treatments at the time of COVID breakdown in the field of internet of things (IoT). In [8], authors performed an analysis on the Markov random fields (MRF) for medical applications. However, in big data manual processing is one of the complex tasks and MRF play a vital role in overcoming medical related issues in big data. In addition, some other works like [9], researchers made a survey on privacy attacks location and creates a prevention deployment on IoT based vehicular application. For addressing the problems due privacy attacks, anonymity and cryptographic interpretations are introduced on the basis of digital signature technique. However, while integrating these techniques into the big data the performance gets very much reduced and lacks its capability.

Nowadays, PPDM (privacy preserving data mining) is a big challenge and is a technique used to transform data to preserve privacy. The growth in data mining leads to novel research as PPDM. There are several techniques used to achieve PPDM. They are anonymization, condensation, Perturbation, Cryptography, and Randomization [10]. Anonymization eliminates the information from the various

information sets to generalize the attributes [11]. Condensation is achieved by forming clusters in several ways. The range of each cluster differs from that of any other cluster [12]. Cryptography is protected with the help of record encryption to find identity and sensitive information [13]. In randomization, the records are shuffled vertically to hide the correct identity. The perturbation indicates that information is being manipulated using noise.

However, various sectors especially in health care, the amount of data gets increased exponentially and it is commonly termed as big data. Perturbing these kinds of data is one of the challenging task due to increasing intruders and advancement in data mining process [14]. Data perturbation is one the privacy preservation approaches that can modify the data present in the database to solve the individual's confidentiality problem. Some of the existing perturbation techniques like random rotation [15], condensation [16] and micro-aggregation [17] remain challenging for balancing both the privacy and the accuracy. In addition, data present in the health care database increases endlessly and intruders have high chance in extracting the information. In recent era, differential privacy (DP) [18] has attracted the researchers a lot by introducing k-anonymity and l-diversity and providing high data securing in data mining.

The noise can be classified into addictive and multiplicative noise [19]. Addictive noise- also called Gaussian noise that can be looked slightly blurry and soft. It is the undesired instantaneous signal and gets added to some real signals. The image of each pixel changes from its real values by a small number. Multiplicative noise is undesired instantaneous signals that get multiplied into real signals. Data perturbation is a method of preserving data and maintaining confidentiality [20]. Several ML models are used to perturb the data and are efficiently used in PPDM. Further, it is compared with the original data to evaluate the efficiency of the model [21] [22]. ML is a computational science field that verifies and interprets the pattern. ML model-based approaches handle big data in the perturbation process [23]. Hence this work used an optimization-based ML model for data perturbation.

A. Motivation

Recent privacy-preserving data mining (PPDM) is a hot research topic in the big data mining process. The major aim of the PPDM is to protect individual data privacy from third party intruders. Nowadays, users are more aware of privacy intrusions on personal data and hesitate to share personal information. Several approaches are present to preserve data like perturbation, anonymization and data transformation. PPDM by data perturbation has attained popularity because these models can hide secret data successfully and ensure more utility in the retrieval process. Some perturbation models are geometric perturbation, condensation, additive perturbation, micro aggregation, and random rotation aids to perturb the medical data effectively. In the medical field, preserving data is essential since medical documents are relevant to privacy concerns and human subjects. PPDM perturbs the data, and intruders cannot identify the medical data. However, many techniques fail to hide the difference between the original and perturbed data, which helps the intruders to easily identify the perturbation in the medical data. Hence, the accuracy of the original and privacy preserved data degrades and cannot apply to the medical data mining process. The latest advancement in the big data mining process helps to achieve better perturbation without the knowledge of intruders.

Recently, numerous approaches have been proposed in the literature for preserving data. But these methods have many limitations. Existing PPDM has high computational complexity for a large volume of data. Further inefficiency poor scalability and make data reconstruction impossible with these approaches. These major drawbacks motivate us to develop an enhanced DL model for medical data mining privacy preservation. Hence, this work proposes the model improved perturbation IGDP (Geometric Data Perturbation) with three-dimensional rotation (3DR) for privacy preservation. The proposed technique is integral in preserving privacy in health care data under low time complexity. *The major contributions of the proposed model are illustrated below:*

- To introduce a novel perturbation technique (IGDP-KSVM-HHO) for privacy preservation in healthcare data
- To pre-process the data using the min-max normalization technique for better data privacy.
- To reduce the data dimensions by introducing hybridized SVD-PCA technique to avoid data redundancy.
- To cluster the health care data using the IFCM technique, the hierarchical data is segmented, and every partition is set as the cluster.
- To perform perturbation using the IGDP technique for the clustered data to improve health care privacy preservation.
- To propose a KSVM-HHO approach for classifying the perturbated health care data accurately.
- To implement the proposed method in PYTHON, performance measures like F-measure, precision, accuracy, execution time and MSE are analyzed and compared with existing techniques.

The remaining section of the research work is arranged as follows: Section II is the recent relevant literature work; Section III depicts the proposed perturbation technique; Section IV gives the overall evaluation of implemented results. Section V provides the conclusion of the work.

II. Related Works

Some of the recent related work based on privacy preserving data mining is listed below:

Devi and Manikandan [24] introduced a rotation based condensation technique with geometric transformation for PPDM. This model provided better resilience over the attacks on the data reconstruction process. The improved P2RoCAl was utilized to measure the dynamics of classification

accuracy. This model provided less time for the empirical process, proving that it could be efficiently used in big data analysis. Kousika and Premalatha [25] proposed SVD (singular value decomposition) and 3RDP (3D rotation data perturbation) for PPDM. Using these models, a perturbed matrix was obtained. Several ML classifiers classified the original and perturbed data, and the evaluation was computed based on the accuracy rate. Experimentation proved that SVD-3RDP outperformed by attaining better accuracy for matrices of various sizes.

Kumar and Premalatha [26] used information value and weight evidence for the initial data perturbation. After that, the 3D shearing was fed on a quasi-identifier once the initial data perturbation was done. Then, several ML classifiers were fed on three benchmark datasets for analyzing the original and perturbed data. The experimentation was analyzed on 2D and 3D rotation. This model can preserve data utility with more privacy preserving and transformation capacity.

Chamikara et al. [27] introduced a novel perturbation model PABIDOT for addressing the utility problem of traditional data perturbation techniques. This model provides privacy to data via Φ – separation. The accuracy outcome of the perturbed data was near the original data. This model attained a systematic model for optimizing data perturbation variables. Finally, privacy and utility were analyzed on certain metrics.

Kumar et al. [28] proposed HER (Electronic Health Records) for hiding sensitive data by integrating fuzzy and association rules. The fuzzification was formed when multilevel privacy approaches were applied, and association rules perturbed outcomes. The experimentation was carried out on the UCI repository, and ML models were used to compute accuracy to show the robustness of the model.

Bedi and Goyal [29] defined privacy preservation in medical data in cloud IoT using extended fully homo-morphic encryption (EFHE). In this study, the FHE helps in adding and multiplying the cipher text. Finally, the information present in the medical data maintains perturbation and shows perturbated results. Thus, the attackers were unaware of the perturbed input and output states during the data mining process. In the experimental section, the peak-to-signal error (PSNR) of 30dB and SNR of 10dB were obtained. However, the security level of this method needs to be increased further for processing big data.

Reddy and Rao [30] defined the clustering and GDP technique for privacy preservation in health care data. The hierarchical data were clustered in this work using the k-means clustering technique. Finally, the clustered data was perturbated on the basis of the GDP algorithm. The perturbed values were encapsulated in the public, and clustered data were encapsulated under the private key. The experimental section obtained an accuracy of 79.58% and an execution time of 161.558s. But this method takes high execution time and obtains the average accuracy.

Janakiraman and Maruthukutty [31] introduced ML based techniques for perturbing the DNA based medical data. In this paper, integrated condensation based PP rotation based DP and ensemble classification (ICS-PPR-DPEC) was emphasized to secure medical data. At the initial stage, condensation algorithm based DP (CADP) was introduced to group the data under tuple distances. Finally, an ensemble machine learning (ELM) based classifier was utilized for recognizing the perturbed human DNA based medical data. In the experimental section, an accuracy of 93.2%, precision of 90%, and recall of 89% were obtained. However, this method faces high complexity during perturbation.

Santhana and Natarajan [32] determined big data analysis for health care data based on clustering and DP algorithm. In this study, an improved FKM (IFKM) based clustering algorithm was introduced to cluster the medical data. Then, modified 3D rotation based DP was introduced to preserve the privacy of the medical data. The experimental section obtained an accuracy of 94% and an execution time of 35ms. However, this method lacks its performance when the data gets increased.

Sujatha and Udayarani [33] evaluated chaotic based geometric DP (CGDP) and hierarchical approach for preserving privacy in health care big data. Initially, the CGDP technique was introduced to perturb the healthcare data. Then, homomorphism based ensemble gradient approach was introduced to classify the perturbed data accurately. In the experimental section, an accuracy of 87% was obtained. However, this technique lacks a clustering technique; hence, the origin of the data cannot be identified.

Even though the approaches mentioned above' outcomes are encouraging, these methods have some limitations. These models take more time to process the data. Further, these methods do not seem to provide full security for the data. Hence, an efficient perturbation model is essential for maintaining the confidentiality of health data.

A. Problem Statement

The recent advancements in the big data mining process have grown much attention towards researchers, especially in the health care sector. With the fast growth in technological advancements, third party and other adversarial attacks also increase exponentially. However, in big data, enormous amounts of data are blemished daily with the increasing data mining process. The third party's utilization of individual privacy details increases for various commercial purposes. Nowadays, data perturbation is one of the hot topics in preserving one's privacy effectively. Many different data perturbation techniques have been introduced for effectively hiding personal details from attackers. But each has its benefits and disadvantages during data perturbation. For big data, recent models lack the capability and face high time complexity. In addition, the existing techniques fail to preserve without the knowledge of intruders, which becomes one of the open issues in many data mining applications.

Nowadays, with the latest technologies, the data can be perturbated without the knowledge of attackers, whether the data is perturbed or original. After the data perturbation, the origin of the data cannot be detected in many recent studies. To the best of our knowledge, the proposed big data mining

method outperforms well in preserving privacy without any knowledge to the intruders.

III. Proposed Methodology

Due to the emergence of ICT (Information and Communication Technology), healthcare data are saved in electronic form and obtained based on the requirements. However, big data privacy determines managing big data under minimal risk and secures the hyper-sensitive data. Generally, big data is spread all over the locations, which destroys patients' privacy for various purposes. Traditional privacy process lacks in handling big data especially in the healthcare sector. Hence, this article privacy enhanced ML algorithms for preserving medical data in the big data mining process. At the initial stage, the min-max normalization based pre-processing technique is emphasized to normalize the medical data efficiently. The normalized data is then fed into the PCA-SVD technique for dimensionality reduction. Then IFCM clusters the data to avoid unwanted complexities while processing big data. In addition, IGDP (Geometric Data Perturbation)-three-dimensional rotation (3DR) is used to effectively preserve data privacy from external attacks. For the classification of perturbed data, an optimization-based kernel Support Vector Machine (KSVM) is utilized. Further, for optimizing the weight of KSVM and improving the performance in classification, a meta-heuristic optimization technique HHO is proposed in this work. Fig. 1 depicts the framework of the introduced Perturbation model.

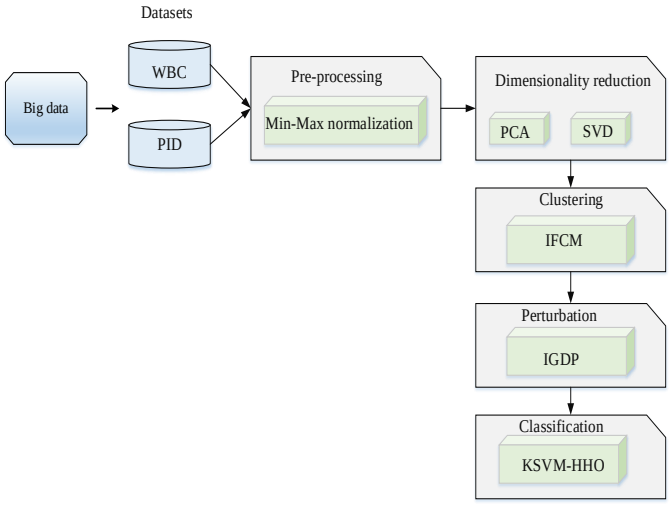

Fig. 1. Framework of the introduced perturbation model.

A. Min-Max Normalization

For the classification of perturbed medical data, pre-processing is the essential stage that can effectively improve the performance accuracy of the proposed technique. In the pre-processing stage, data normalization is undertaken to minimize data redundancy and error during perturbation. Recently studies have failed to normalize the data and consume high time complexity while perturbing the medical data. The proposed min-max normalization technique aids in normalizing the data and efficiently preserving the dataset's sensitive attributes. The major aim of the proposed model is to normalize the original dataset E into a preserved dataset $E^{'}$ that satisfies the needs of privacy with better data privacy. In the dataset, every attribute is normalized arranging its value, hence they will come to the range of 0-1. To map the value $D^{'}$ of the attribute b from the range $[\max_b, \min_b]$ to $[New_{\max_b}, New_{\min_b}]$ is given by:

$$D^{'} = \frac{D - \min_b}{\max_b - \min_b}(New_{\max_b} - New_{\min_b}) + New_{\min_b} \quad (1)$$

Where D and $D^{'}$ are the original and newly computed value. $\max_b$ and $\min_b$ are the attribute's maximum and minimum values. $New_{\max_b}$ and $New_{\min_b}$ are the attribute's new maximum and minimum values.

B. Dimensionality Reduction

After pre-processing, the dimensionality of the data is reduced for the normalized data. Similar features present in the medical data are completely removed from the medical data for the minimization of execution time. This research introduces a hybrid PCA-SVD technique for the elimination of similar features. The major aim of PCA is to reduce the dimension of the training set, and the major component of the training set is the outcome of the projection of the Eigenvector with Eigenvalue. The procedure of dimensionality reduction using PCA is given below:

Let $y \in S^n$ is a test sample with n parameters, and every parameter has m independent samples. Then the data matrix is written as:

$$Y_n = [y(1), y(2), ..., y(m), \quad Y_n \in S^{n \times m} \quad (2)$$

Every column of Y_n indicates a parameter, and every row indicates the sample. Since the dimensions of the measured parameters are different, every column of data is normalized, and it is represented as:

$$Y^* = \frac{Y_n - d.p}{diag\ (\sigma)} \quad (3)$$

Where $d = (1,1,....1)^T$, p is the mean of Y columns and $diag\ (\sigma)$ is a parameter matrix of Y_n.

The covariance matrix of Y^* is:

$$V = \frac{1}{m-1} Y^{*T} Y^* \quad (4)$$

The matrix processing is generally the decomposition of the Eigen value. The matrix is sorted in descending order based on the Eigenvalues size. The value of Y^* is decomposed by:

$$Y^* = \vec{Y}^* + F = TR^T + F \quad (5)$$

$$T = Y^* R \quad (6)$$

Where $\vec{Y}^*$ is a projection in PCA, residual space of projection is F , and load matrix is denoted as R. T is a scoring matrix and the components in T is a primary parameter. PCA is a modelling segment, residual space is not a modelling segment, and it depicts the noise. The principal component number is selected according to CPV (Cumulative Percent Variance). This scheme represents the number of principal components based on cumulative principal elements. CPV is a ratio of the data variation defined by the initial principal component to the total data variation. Hence, CPV is represented as:

$$CPV = \frac{\sum_{j=1}^{A} \gamma_j}{\sum_{j=1}^{m} \gamma_j} \quad (7)$$

Where γ_j is an Eigenvalue of V. The features reduced by PCA are given to SVD for further dimensionality reduction. SVD is used for eliminating correlation between data features. In SVD, each sample is perturbed using the same parameter. Let a matrix indicates the original data B with dimension $l \times m$. The column indicates the attributes in the matrix, and the row indicates data objects. The SVD of the matrix C is given in equation (8).

$$C = XBY^T \quad (8)$$

Where X and Y^T are the $l \times l$ and $m \times m$ orthogonal matrix and B is $l \times m$ diagonal matrix.

The decomposition matrix C in (8) is represented as:

$$C = \sum_{j=1}^{r} \sigma_j B_j Y_j^T \quad (9)$$

Where σ_j is a singular value of C and the columns of X is B_j and Y_j. The SVD of the matrix C is used for solving the linear model $Cy = d$, and it is given by:

$$y^+ = \sum_{j=1}^{r} \sigma_j^{-1} < d, B_j > Y_j^T \quad (10)$$

C. Clustering using IFCM

The dimensionality reduced data is then fed into the IFCM clustering technique to cluster the medical data having similar fields. In the traditional FCM [34] technique, the cluster's centre and the cluster numbers are fixed artificially, which is sensitive for the first cluster centre. Further, this algorithm has slow convergence and less stability. Hence, IFCM is introduced, in these clusters, centres are based on two parameters like distance (d_j) and local density (ρ_j). Then the distance of density is given as:

$$\phi_j = d_j \rho_j \quad (11)$$

Then these parameters are fed to FCM to obtain the clustering results. Consider $Y = \{Y_1, Y_2 .... Y_n\}$ is a collection of m is number of clusters and this m is divided into G fuzzy groups. Then the centre matrix is given as $U = \{U_1, U_2 .... U_G\}$, and the objective function to define FCM is given by:

$$K(W,U) = \sum_{j=1}^{G} \sum_{k=1}^{m} (w_{jk})^n (d_{jk})^2 \quad (12)$$

Where W is a dimensional membership matrix, w_{jk} is a membership among Y , and U. d_{jk} is a Euclidean distance among k^{th} sample and j^{th} cluster centre. This equation (12) should satisfy the below conditions.

$$\begin{cases} \sum_{j=1}^{G} w_{jk} = 1, \quad k = 1,2,.....m \\ 0 \le w_{jk} \le 1, \quad j = 1,2,.....G; k = 1,2,.....m \\ 0 < \sum_{k=1}^{m} w_{jk} < m, \quad k = 1,2,.....G \end{cases} \quad (13)$$

Finally, the centre of the cluster is identified by the following expression

$$U_{jk} = \frac{\sum_{k=1}^{m} (w_{jk})^n y_k}{\sum_{k=1}^{m} (w_{jk})^n} \quad (14)$$

Where, y_k is a data in cluster and the membership function will be updated based on centres of the cluster, and it is represented as:

$$M_{jk} = \frac{1}{\sum_{l=1}^{m} \left(\frac{\| y_k - U_{jk} \|}{\| y_k - U_{kl} \|} \right)^{2/(n-1)}} \quad (15)$$

According to Equations (14) and (15), the dataset is grouped, and during clustering, the data are shuffled into several groups. This clustering is used for performing data perturbation.

D. Improved Geometric Data Perturbation (IGDP)

After clustering, the data perturbation process is undertaken to provide privacy to the medical data. Recently many perturbation techniques have been introduced to preserve medical data effectively without the intruder's knowledge. However, due to high processing time, those techniques cannot be applicable for processing huge volumes of data, especially in big data. The perturbation technique IGDP is used to perturb the clustered data. 3DR is used with GDP, and it is used to distort data by rotating three orientations (S_x, S_y, S_z) , and axes pairs utilized for the rotation are (S_{xy}, S_{yz}, S_{zx}). The rotation operation is provided more than once until the entire attributes are transformed to preserve privacy. The procedure of 3DR is given below:

Stage 1: Choose the three orientations (S_x, S_y, S_z) and compute the rotation matrix as:

$$S_{xy} = S_x \times S_y = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ \sin^2\theta & \cos\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & -\sin\theta & \cos^2\theta \end{pmatrix}$$

$$S_{yz} = S_y \times S_z = \begin{pmatrix} \cos^2\theta & -\sin\theta\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta & 0 \\ \sin\theta\cos\theta & -\sin^2\theta & \cos\theta \end{pmatrix}$$

$$S_{zx} = S_z \times S_x = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta\cos\theta & \cos^2\theta & \sin\theta \\ -\sin^2\theta & \sin\theta\cos\theta & \cos\theta \end{pmatrix}$$

Stage 2: The attributes are grouped into three functions (A_x, A_y, A_z).

Stage 3: Three functions are rotated around (S_x, S_y, S_z) in a 3D plane for several rotation angles.

Stage 4: Identify the rotation angle until all attributes are transferred to preserve privacy.

GDP has a sequence of randomized geometric transformations like multiplicative transformation (M), distance perturbation (Δ) and translation transformation (φ).

$$g(y) = My + \Delta + \varphi \quad (16)$$

In this, the element M_T is a rotational matrix and y is original data.

1) Multiplicative transformation: The parameter (M) is a rotation matrix. This matrix accurately preserves distance. Let the rotation perturbation is expressed as:

$$g(y) = My \quad (17)$$

The orthogonal matrix is given as $M_{d\times d}$ and has some characteristics. The transpose of M is M^T, the Identity matrix is I_j and m_{jk} is the (j,k) component of M. The matrix of M in columns and rows is orthonormal. The resultant matrix is also orthonormal when the orders of columns and rows are changed.

2) Translation transformation: This transformation is written as:

$$g(y) = y + \varphi \quad (18)$$

Let us consider two points a and b in the original space, and the distance is given as:

$$\| (a-t)-(b-t) \| = \| a-a \| \quad (19)$$

The translation saves distance and doesn't protect the inner product. When translation and rotation are integrated, φ can enhance the preservation.

3) Distance perturbation: The major aim of this perturbation is to preserve distance and provide strength to distance inference attacks. The entry of the random matrix $\Delta_{f\times m}$ is an independent sample exhausted from the same distribution with less variance and zero means. When this random matrix is included, the distance between the point's pair gets disturbed. This IGDP provides more security to medical data. Then fully obtained perturbed matrix is applied for the classification process.

E. Classification using KSVM-HHO

The Perturbed data is then fed into the classification stage to classify the perturbed data accurately. The Classification stage helps to assess a dataset that retains data mining performance approaches after data perturbation. Recently, many ML classifiers have been utilized as a classifier for perturbed data, and each has its benefits and drawbacks. In this work, the ML based SVM classifier is used for classifying the perturbed data and utilizes a separation of non-linear mapping to transform original trained data, which are linearly transformed by a kernel function. The Polynomial kernel function (PKF) is used for the transformation, and this function transforms the non-linear into high dimensionality features. Hence, the data partition is feasible, making a classification task more convenient. Normal SVM states a hyperplane that is separated into two training classes, and it is expressed as:

$$g(z) = Y_k^* \phi_k (z) + f \quad (20)$$

where, Y_k^* is a hyper-plane parameter, f is a bias value and $\phi_k(z)$ is a term used for mapping vector z into high dimensionality space. The major aim is to make an effective classifier by the training set (o_k, z_k). The optimal value of Y_k^* and f is given by:

$$o_k(Y_k^* z_k + f) \geq 1 - \zeta_k \quad for\ k = 1,2,.....M \tag{21}$$

where, ζ_k is a slack parameter for all k. Y_k^* and ζ_k reduces the cost function, and it is expressed as:

$$\phi(Y_k^*, \zeta_k) = \frac{1}{2} \| Y_k^* \|^2 + R_e \sum_{k=1}^{M} \zeta_k \tag{22}$$

where, R_e is a regularized variable and used to control the size of the discriminant function. The final result of the SVM is denoted as:

$$o_k(y) = \sum_{k=1}^{M} J(y_k, y) \tag{23}$$

where, $J(y_k, y)$ is a kernel function. There are several types of kernels. In this work, a polynomial kernel is utilized. The kernel's parameters are to be adjusted previously to the process of the training data. It is one of the non-stationary kernels, and it works better for normalizing training data. It is denoted as:

$$J(y_k, y) = \gamma((y_k, y) + 1)^d \tag{24}$$

where, d is a polynomial degree. However, the proposed classifier high affected due to increased losses in which the origin of the data cannot be detected. The loss function of the proposed KSVM classifier can be interpreted as,

$$L = \min(Accuracy), \max(MSE) \tag{25}$$

1) Parameter tuning using HHO optimizer: The proposed KSVM classifier degrades its performance while processing huge medical data. Thus, the accuracy performance is much reduced, which can be overcome by tuning the parameters in the classifier model. Recently, many optimization techniques have played an integral role in tuning the parameters, thus enhancing the system's efficiency. This research introduces HHO based metaheuristic optimization technique for parameter tuning of the proposed classifier model.

Initially, the parameters of HHO are initialized, and the parameters of the polynomial SVM are encoded and trained. *Obtain* fitness value and classification of horses based on age. Then, the position of the horse is obtained. The process is repeated until the satisfied criteria are met.

This optimization is based on the horse's behaviour. There are six patterns of behaviours they are Grazing (G), Hierarchy (H), Roaming (R), Sociability (S), defences (D) and Imitation (I). The following equation is based on a movement provided to horses at every iteration.

$$Z_n^{i,age} = \vec{U}_n^{i,age} + Z_n^{(i-1),age} \tag{26}$$

Where $Z_n^{i,age}$ is a n^{th} horse position, age is the age of the horses, i is a present iteration and $\vec{U}_n^{i,age}$ is a velocity vector. The following steps show the horse's six patterns of behaviour.

a) Grazing (G): Horses feed on grasses, grains and plants. HHO creates the grazing field around every horse with a factor k. Horses graze at any age in their entire lifetime. The grazing behaviour is expressed as:

$$\vec{G}_n^{i,age} = g_i(LB + \rho \times UB) + [Z_n^{(i-1)}] \tag{27}$$

$$g_n^{i,age} = g_n^{(i-1),age} \times \sigma_g \tag{28}$$

Where $\vec{G}_n^{i,age}$ is a parameter of motion, and it shows the ability of horse grazing tendency. This term minimizes linearity by σ_g. LB and UB are the lower and upper bounds, which range from 0 to 1.

b) Hierarchy (H): It is proved that the horses at 5 to 15 years are used to follow hierarchy rules, and it is indicated as:

$$\vec{H}_n^{i,age} = h_n^{(i-1),age}[Z_*^{(i-1)} - Z_n^{(i-1)}] \tag{29}$$

$$h_n^{i,age} = h_n^{(i-1),age} \times \sigma_h \tag{30}$$

Where $h_n^{i,age}$ is the effect of the best location of the horse on the velocity, σ_h is a reducing factor and $Z_*^{(i-1)}$ is the best horse location.

c) Sociability (S): This behaviour is regarded as the movement to an average position of other horses. Horses in their middle age have an interest in the herd, and it is expressed as:

$$\vec{S}_n^{i,age} = s_n^{(i,age)}\left[\left(\frac{1}{N}\sum_{l=1}^{N} Z_l^{(i-1)}\right) - Z_n^{(i-1)}\right] \tag{31}$$

$$s_n^{i,age} = s_n^{(i-1),age} \times \sigma_s \tag{32}$$

Where $\vec{S}_n^{i,age}$ is a social movement vector, σ_s is a reducing factor and $S_n^{(i,age)}$ is an orientation of a horse to a herd in i^{th} iteration.

d) Imitation (I): This characteristic of a horse is set as i, and the horse in 0 to 5 years tries to mimic other horses. This imitating behaviour is expressed as:

$$\vec{I}_n^{i,age} = i_n^{(i,age)}\left[\left(\frac{1}{qN}\sum_{l=1}^{qN}\vec{Z}_l^{(i-1)}\right) - Z_n^{(i-1)}\right] \quad (33)$$

$$i_n^{i,age} = i_n^{(i-1),age} \times \sigma_i \quad (34)$$

Where $\vec{I}_n^{i,age}$ movement vector of the horse to the average of best horses with a position of $\vec{Z}$. The total of horses with the best location is qN and σ_i is a reducing factor.

e) Defence (D): This behaviour exists in their overall lifetime. This horse mechanism is represented as d, and it is a negative factor in equations (32) and (33).

$$\vec{D}_n^{i,age} = -d_n^{(i,age)}\left[\left(\frac{1}{sN}\sum_{l=1}^{qN}\vec{Z}_l^{(i-1)}\right) - Z_n^{(i-1)}\right] \quad (35)$$

$$d_n^{i,age} = d_n^{(i-1),age} \times \sigma_d \quad (36)$$

Where $\vec{D}_n^{i,age}$ is an escaping vector from a bad location $\vec{Z}_l$, sN is a total of horses and σ_d is a reducing factor.

f) Roaming (R): This characteristic is imitated as a random motion and represented using r. This behaviour is generally seen at 0 to 5 years and goes away at middle age. It is represented as:

$$\vec{R}_n^{i,age} = r_n^{(i,age)}\ \vec{qZ}_l^{(i-1)} \quad (37)$$

$$r_n^{i,age} = r_n^{(i-1),age} \times \sigma_r \quad (38)$$

Where $\vec{R}_n^{i,age}$ is a random velocity vector and σ_r is a reducing factor of $r_n^{i,age}$.

g) Fitness function (F): Finally, the parameter is tuned, and its weight is updated based on the updated location of the proposed optimizer. The fitness function can be formulated as,

$$fitness\ \ function(f) = \max(Accuracy), \min(MSE) \quad (39)$$

As shown above, the data is perturbed using IDGP and classified by KSVM-HHO. Several classifiers classify perturbed and original data; the achievements are given in the following section. Fig. 2 illustrates the flowchart of the proposed IGDP-KSVM-HHO technique. Algorithm 1 depicts the pseudo-code for the proposed method.

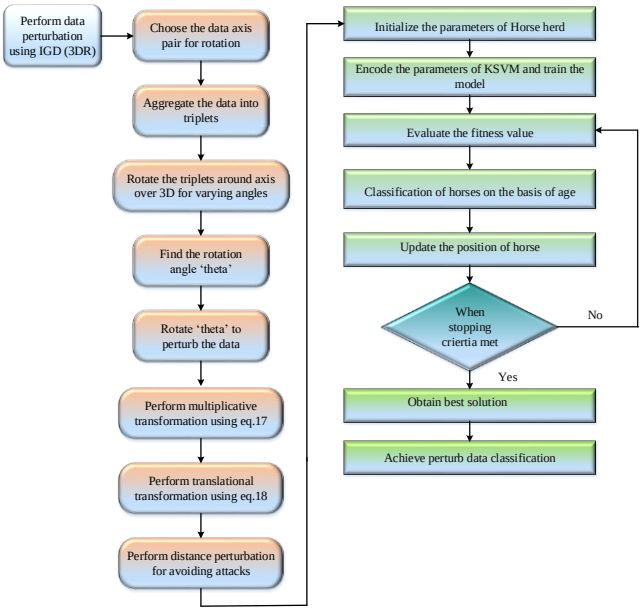

Fig. 2. Flowchart of the proposed IGDP-KSVM-HHO technique.

Algorithm 1: Pseudo code for the proposed method

Pseudo code for the proposed method

Input: Noisy medical dataset from the raw dataset;
Output: Classified perturbed and original medical data;
Start data pre-processing using min-max normalization
Perform data normalization and preserve sensitive attributes using equation 1; //normalized medical data
For dimensionality reduction
Utilize hybrid PCA-SVD technique;
Consider the data as a data matrix of PCA using equation 2;
Normalize the medical data present in the row and columns using equation 3;
Calculate the covariance matrix V using equation 4;
Arrange the matrix into descending order and measure the score matrix using equations 5 and 6;
Analyze the data variation CPV in the matrix using equation 7;
While the correlation between the features is high
Integrate SVD with PCA for further dimensionality reduction;
Consider the data matrix of SVD using equation 8;
Calculate decomposition matric C using equation 9;
Calculate the dimensionality reduced matrix using equation 10; //dimensionality reduced medical data
End while
End for
Do clustering for the dimensionality reduced medical data

using IFCM //helps to find origin of data

Calculate density distance for determining the cluster centre using equation 11;

Assume the centre matrix $U = \{U_1, U_2 .... U_G\}$ having random clusters;

Calculate the objective function using equation 12;

Assume the condition for becoming a member of a particular cluster;

If $(k = 1)$

Calculate the membership matrix M_{jk} having cluster centre using equation 15;

Update the cluster centre U_{jk} using equation 14;

$k \leftarrow K + 1$

Return best clusters having membership function M

Perform data perturbation using IGDP (3D rotation);

Input: Clustered medical data, its attributes and array of security thresholds;

Output: Perturbed medical data;

For GDP with 3D rotational transformation

Select the orientations as, (S_x, S_y, S_z) and perform rotation as, S_{xy}, S_{yz} and S_{zx};

Group the attributes as, (A_x, A_y, A_z);

Rotate (S_x, S_y, S_z) in a 3D plane for varying angles;

Do geometric transformations to preserve distance privacy and avoid attacks

Perform multiplicative transformation using equation 17;

Perform translation transformation using equation 18;

Preserve distance and eliminate attacks using distance perturbation;

End for

Return perturbed medical data

For classification of data perturbation

Do optimized KSVM-HHO for perturbed data classification

Initialize the parameters of HHO using equation 26;

Train the SVM model using equation 20;

Calculate the fitness value for the reduced cost function using equation 22;

Calculate the age based on the hierarchical rule in equations 28 and 29;

Find the optimal value Y_k^* and f by tuning the parameters using equation 21;

$Y_k \leftarrow Y_k^*$;

Analyze the data obtained by optimized SVM using equation 23;

Adjust the kernel parameters using equation 24;

Update the horse position using equations 27 and 28;

Analyze the velocity of the horse based on the age using equations 37 and 38;

If fitness condition is satisfied

Generate best outcome;

Else;

Repeat $T \leftarrow T + 1$

End if

Return the accurate classified perturbed medical data

Stop

IV. Results and Discussion

The proposed IGDP-KSVM-HHO is evaluated in the Python platform. The efficiency of the approach is evaluated based on accuracy, F-measure, MSE, precision, sensitivity, specificity and execution time. The performance of Perturbed and original data are classified by ML classifiers like KSVM, SVM, NB (Naïve Bayes), KNN (K nearest neighbour) and RF (Random forest). Table I depicts the simulation parameters of the proposed method. Table II represents the system configuration of the proposed method.

TABLE I. Simulation Parameters of the Proposed Method

Simulation Parameters	Values
Optimizer	HHO optimizer
SVM-type	C-classification
SVM-kernel	Polynomial kernel
Cost	1
Gamma (γ)	0.0625
Number of support vectors	8434

TABLE II. System Configuration of the Proposed Method

System Specifications		
S. No	**Parameter**	**Configuration**
1	Device name	Desktop
2	Processor	Intel(R) Core(TM) i3-7100 CPU @ 3.90GHz, 3912 MHz, 2 Core(s), 2 Logical Processor(s)
3	Installed RAM	8.00 GB (7.89 GB usable)
4	Device ID	DFBCDAE4-A190-457D-8C56-FDDDBB348B4F
5	Product ID	00330-50186-83065-AAOEM
6	Pen and Touch	No pen or touch input is available for this display
7	System Type	64-bit operating system, x64-based processor

A. Dataset Detail

1) WBC dataset: This dataset is obtained from the UCI machine learning repository and used to record Brest cancer cases' measurements. It has nine attributes and 699 samples. This dataset has two classes, and it is downloaded using the following URL. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(Prognostic)

2) PID dataset: This dataset is initially obtained from the National Institute of kidney, digestive and diabetes diseases. It has eight attributes and 768 samples. This dataset is downloaded using the https://www.kaggle.com/uciml/pima-indians-diabetes-database/version/1. For experimental analysis, both datasets are divided into 70% for training and 30% for testing.

B. Performance Measures

In order to quantitatively calculate the performance of the developed scheme, certain metrics are utilized. This research uses six metrics, execution time, sensitivity, accuracy, F1 score, and precision, to evaluate the performance. The introduced IGDP-KSVM-HHO was evaluated based on True positive (T_p), false positive (F_p), true negative (T_n) and false negative (F_n). The description of all metrics with the formula is described below.

1) Accuracy (A)*:* It is a ratio of the number of exact predictions to the overall prediction. It is expressed as:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{40}$$

2) Sensitivity (Se)*:* An amount of active positive is exactly found as positive using the classifier. The following expression represents it:

$$Se = \frac{T_p}{T_p + F_n} \tag{41}$$

3) Precision (P)*:* It is a ratio of the positively predicted sample which is positive to the overall observation, and it is expressed as:

$$P = \frac{T_p}{T_p + F_p} \tag{42}$$

4) F-measure (F)*:* It is a harmonic mean of Se and P. The following expression expresses it:

$$F = 2 \times \frac{P \times Se}{P + Se} \tag{43}$$

5) MSE (Mean Square Error): It is measured using the average squared intensity of original and perturbed values. It is denoted as:

$$MSE = \frac{1}{w_i h_i} \sum_{j=1}^{w_i} \sum_{k=1}^{h_i} (D_{jk} - S_{jk})^2 \tag{44}$$

Where D_{jk} and S_{jk} are the grey values of (j,k).

C. Performance using the WBC Dataset

This section compares the performance of IGDP-KSVM-HHO with several ML classifiers on the WBC dataset. The metrics like accuracy, F-measure, sensitivity, precision, MSE and execution time are computed.

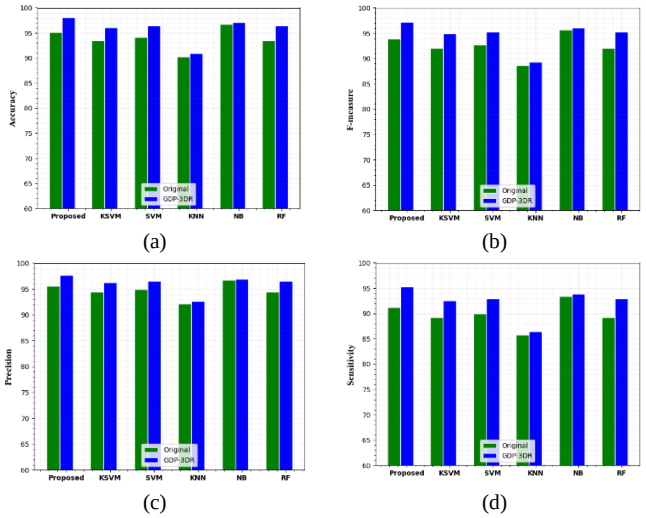

Fig. 3. Comparison of (a) accuracy, (b) F-measure, (c) precision, (d) sensitivity.

Fig. 3 compares several metrics like Accuracy, F-measure, Precision and sensitivity. The performance is carried out for the original dataset and the perturbed dataset. That is, the WBC dataset is perturbed using IGDP. The analysis proved that the proposed IGDP-KSVM-HHO attained better results than existing classifiers. It is seen from the graph that the proposed IGDP-KSVM-HHO attained almost equal to the original dataset. The accuracy attained by the original dataset is 95.11%, and IGDP-KSVM-HHO attained an accuracy of 98.08%, respectively. The proposed model attained better results due to the KSVM optimized by HHO. It shows that this model can provide privacy to data efficiently.

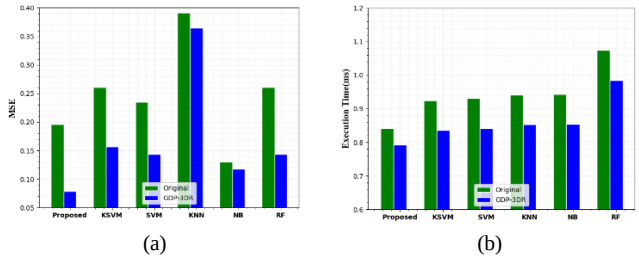

Fig. 4. Comparison of (a) MSE (b) Execution time.

Fig. 4 indicates the comparison of the measures like Execution time and MSE. Execution time is the overall time taken by the proposed model to achieve the outcomes, and it is represented in milliseconds (ms). The proposed IGDP-KSVM-HHO takes only 0.79 ms to complete the process. But, other classifiers take more time to complete the process. It shows that the proposed model has less computational complexity.

Further, the perturbed model's MSE of IGDP-KSVM-HHO, KSVM, SVM, KNN, NB and RF are 0.077, 0.155, 0.14, 0.36, 0.11 and 0.14, respectively. It is proved that the system has fewer errors. This perturbation can be applied to various datasets and utilized for big data analysis.

D. Performance using PID Dataset

This section illustrates the quantitative performance of various classifiers on the PID dataset. The original and

perturbed data is evaluated on a PID dataset, and some performance measures are computed.

Table III depicts the performance comparison of various classifiers. Several classifiers are verified on the perturbation method to evaluate the data perturbation effect in the potential of data mining. The experimental demonstration proved that the introduced model preserves the data and enhances accuracy in classification. The existing models attained poor results due to their structure and high computational complexity. In all cases, the performance attained by the original dataset is nearly equal to the perturbed PID dataset. The accuracy, F-measure and precision obtained by the perturbed dataset for the proposed model are 98.04%, 97.36% and 96.6%, respectively.

Table IV depicts the performance comparison of sensitivity, MSE and Execution time of classifiers like KSVM, SVM, NB, KNN, RF and proposed IGDP-KSVM-HHO. Classifiers are applied for original and perturbed data. The table shows that the overall values of a perturbed dataset are almost the same as the perturbed data. It shows that IGDP-KSVM-HHO ensures better accuracy than existing classifiers. In addition, after perturbation, the mined remains similar, preserving data utility. The proposed model achieved better due to the 3D rotation and optimal weight selection.

TABLE III. PERFORMANCE COMPARISON (ACCURACY, F-MEASURE AND PRECISION) OF VARIOUS CLASSIFIERS

Original dataset				Perturbed dataset		
Classifiers	**A**	**F**	**p**	**A**	**F**	**p**
RF	96.09	95.3	95.55	97.71	96.9	95.78
KNN	96.4	95.64	95.6	96.4	95.5	94.9
NB	91.85	90.95	92.06	94.34	95.63	95.39
SVM	97.39	96.6	96.15	97.71	96.9	95.99
KSVM	97.39	96.6	96.15	97.39	96.66	96.1
IGDP-KSVM-HHO (proposed)	97.06	96.26	95.07	98.04	97.36	96.6

TABLE IV. PERFORMANCE COMPARISON (SENSITIVITY, MSE AND EXECUTION TIME) OF VARIOUS CLASSIFIER

Original dataset				Perturbed dataset		
Classifiers	**Se**	**MSE**	**Execution time (ms)**	**Se**	**MSE**	**Execution time (ms)**
RF	93.70	0.039	33.4	96.79	0.022	1.593
KNN	94.2	0.035	32.85	94.67	0.03	1.437
NB	89.06	0.081	32.86	94.34	0.035	1.437
SVM	95.6	0.026	32.7	96.54	0.022	1.34
KSVM	95.6	0.02	32.7	95.65	0.026	1.34
IGDP-KSVM-HHO (proposed)	97.06	0.02	2.2	96.5	0.019	0.84

E. Performance Comparison under different Techniques

In this section, the performance of the proposed method is compared with different existing techniques and proves that the introduced model is highly efficient and accurate. Some of the existing clustering techniques like fuzzy-c means clustering (FCM), k-means (KM), k-medoids and density based spatial clustering of applications with noise (DBSCAN) are utilized. In addition, some of the existing optimization techniques like a genetic algorithm (GA), particle swarm optimization (PSO), differential evaluation (DE) and monarch butterfly optimizer (MBO) are utilized and prove that the proposed parameter tuning optimizer is better. Table V tabulates the comparative performance under different clustering techniques. Table VI and VII illustrates the performance comparison for different optimization techniques under PID and WBC datasets.

TABLE V. OVERALL ACCURACY PERFORMANCE UNDER DIFFERENCE CLUSTERING APPROACHES

Clustering methods	**Accuracy performance**
Proposed	**97.07%**
FCM [35]	89%
KM [36]	93%
k-medoids [37]	92.2%
DBSCAN [38]	94%

TABLE VI. COMPARATIVE PERFORMANCE UNDER DIFFERENT OPTIMIZATION TECHNIQUES FOR THE PID DATASET

Methods	**Original Dataset**		**Perturbed Dataset**	
	Accuracy	**MSE**	**Accuracy**	**MSE**
KSVM-HHO (Proposed)	**95.11**	**0.79**	**98.08**	**0.077**
KSVM-GA	94.85	0.84	97.62	0.085
KSVM-PSO	94.53	0.89	97.24	0.088
KSVM-DE	94.24	0.93	96.85	0.094
KSVM-MBO	93.83	0.97	95.27	0.097

TABLE VII. COMPARATIVE PERFORMANCE UNDER DIFFERENT OPTIMIZATION TECHNIQUES FOR THE WBC DATASET

Methods	**Original Dataset**		**Perturbed Dataset**	
	Accuracy	**MSE**	**Accuracy**	**MSE**
KSVM-HHO (Proposed)	**97.06**	**0.002**	**98.04**	**0.019**
KSVM-GA	96.97	0.0024	97.42	0.023
KSVM-PSO	96.61	0.0029	97.16	0.027
KSVM-DE	96.29	0.0035	96.74	0.031
KSVM-MBO	95.91	0.0041	96.12	0.037

F. Analysis of the Proposed Method

In this section, the outcome of the proposed method is analyzed under original medical data and perturbed medical data. Tables VIII to XI demonstrate the original database, clustered medical data, 3D rotated medical data, and classified perturbed data. In Table IX, C_1, C_2 and C_3 depict the cluster groups, respectively.

TABLE VIII. ORIGINAL SAMPLE MEDICAL DATABASE

Blood pressure (BP)	Age	Gender	Weight (Kg)	Type of Disease
122	22	Male	75	Diabetes
90	25	Male	82	Vision loss
85	43	Female	55	Inflammatory breast cancer
104	55	Male	69	Kidney failure
113	40	Female	45	Ductal carcinoma
82	32	Female	65	Benign tumour
70	27	Male	72	Insulin malfunctions

TABLE IX. MEDICAL DATA CLUSTERING USING IFCM TECHNIQUE

S. No	Cluster groups	Blood pressure (BP)	Age	Gender	Weight (Kg)	Type of disease
1	C_1	122	22	Male	75	Diabetes
2		90	25	Male	82	Vision loss
3		70	27	Male	72	Insulin malfunctions
4	C_2	104	55	Male	69	Kidney failure
5	C_3	113	40	Female	45	Ductal carcinoma
6		82	32	Female	65	Benign tumour
7		85	43	Female	55	Inflammatory breast cancer

TABLE X. 3D ROTATIONAL TRANSFORMATION VALUES

Blood pressure (BP)	Age	Gender	Weight (Kg)	Type of disease
1024	12005	Male	234	Diabetes
2215	60992	Male	709	Vision loss
7505	22951	Male	550	Insulin malfunctions
3378	72950	Male	988	Kidney failure
9045	87657	Female	1012	Ductal carcinoma
4055	30406	Female	2044	Benign tumour
6650	56987	Female	946	Inflammatory breast cancer

TABLE XI. CLASSIFIED PERTURBED MEDICAL DATA USING IGDP ALGORITHM

Blood pressure (BP)	Age	Gender	Weight (Kg)	Type of disease
1030	12010	Male	240	Diabetes
2225	60998	Male	718	Vision loss
7510	22961	Male	550	Insulin malfunctions
3384	72970	Male	999	Kidney failure
9050	87664	Female	1025	Ductal carcinoma
4066	30417	Female	2050	Benign tumour
6659	56998	Female	968	Inflammatory breast cancer

V. CONCLUSION

Due to the advancement of technology, several medical data are frequently gathered and delivered to the institution. Several resources are involved in data collection, analysis and sharing the data, leading to an increase in concerns regarding patients' data. The PPDM model provides several techniques for preserving the data. This paper aims to provide privacy to medical data using the perturbation technique. Two benchmark datasets are selected for this purpose. Initially, the dataset is pre-processed, and the dimensionality is reduced. Then, the reduced features are clustered using IFCM. This clustered data is perturbed by IGDP, which integrates GDP and 3DR. Finally, the perturbed data is classified by the KSVM-HHO classifier. The performance of IGDP- KSVM-HHO is compared to the other ML classifiers like KSVM, SVM, NB, KNN and RF. The performance obtained by IGDP-KSVM-HHO is superior to other models. Moreover, the classification performance of original and perturbed data is almost equal, showing that this model can provide better privacy. For the WBC dataset, the proposed method obtains an overall accuracy of 95.11% and 98.08% for original and perturbation in data. For the PID dataset, the proposed method obtains an overall accuracy of 97.06% and 98.04%, respectively. However, security is the major concern for the big data mining process due to the increase in harmful intruders. The advantage of the proposed method is that it is one of the highly recommended systems for preserving the individual's privacy data under low complexity. Despite this, the proposed method suffers due to high granular access control and faces complexity in detecting the origin of data. In the future, researchers need to focus on developing secure encryption and trust computing techniques to maintain the balance between security and the efficiency of the data mining process. In addition, the researchers need to process the proposed work with various other fields like banking, military sectors etc. and analyze the performance of the same.

REFERENCES

[1] D.F. Sittig, & H. Singh, "A new socio-technical model for studying health information technology in complex adaptive healthcare systems," In Cognitive informatics for biomedicine Springer, Cham, pp. 59-80, 2015.

[2] O. Turel, A. Romashkin, & K.M. Morrison, "Health outcomes of information system use lifestyles among adolescents: videogame addiction, sleep curtailment and cardio-metabolic deficiencies," PloS one, Vol. 11, no. 5, pp. e0154764, 2016.

[3] T. Patel, & V. Patel, "Data privacy in construction industry by privacy-preserving data mining (PPDM) approach," Asian Journal of Civil Engineering, Vol. 21, no. 3, pp. 505-515, 2020.

[4] A. Idri & I. Kadi, "A data mining-based approach for cardiovascular dysautonomias diagnosis and treatment," In 2017 IEEE International Conference on Computer and Information Technology (CIT) IEEE, pp. 245-252, 2017.

[5] J. Liu, Y. Tian, Y. Zhou, Y. Xiao, & N. Ansari, "Privacy preserving distributed data mining based on secure multi-party computation," Computer Communications, Vol. 153, pp. 208-216, 2020.

[6] M.K. Hasan, S. Islam, I. Memon, A.F. Ismail, S. Abdullah, A.K. Budati, and N.S. Nafi, "A novel resource oriented DMA framework for internet of medical things devices in 5G network," IEEE Transactions on Industrial Informatics, Vol. 18, no. 12, pp. 8895-8904, 2022.

[7] A. Khan, J.P. Li, F. Hasan, I. Memon, and A.U. Haq, "Toward analyzing the impact of healthcare treatments in industry 4.0 environment—a self-

care case study during COVID-19 outbreak," In Data Science for COVID-19, pp. 243-256, 2022. Academic Press.

[8] R.A. Shaikh, J. Li, A. Khan, and I. Memon, "Biomedical image processing and analysis using Markov random fields," In 2015 12th International computer conference on wavelet active media technology and information processing (ICCWAMTIP), pp. 179-183, 2015. IEEE.

[9] N. Ahmed, Z. Deng, I. Memon, F. Hassan, H.K. Mohammadani, and R. Iqbal, "A Survey on Location Privacy Attacks and Prevention Deployed with IoT in Vehicular Networks," Wireless Communications and Mobile Computing. Vol. 2022, 2022.

[10] R. Ratra, & P. Gulia, "Privacy preserving data mining: Techniques and algorithms," SSRG International Journal of Engineering Trends and Technology, Vol. 68, no. 11, pp. 56-62, 2020.

[11] A. Majeed, & S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," IEEE Access, Vol. 18, pp. 8512-45, 2020.

[12] M. Rafiei, & W.M. van der Aalst, "Privacy-preserving data publishing in process mining," In International Conference on Business Process Management, pp. 122-138, 2020. Springer, Cham.

[13] M.H. Gerards, C. McCrum, A. Mansfield, & K. Meijer, "Perturbation-based balance training for falls reduction among older adults: Current evidence and implications for clinical practice," Geriatrics & gerontology international, Vol. 17, no. 12, pp. 2294-2303, 2017.

[14] S. Upadhyay, C. Sharma, P. Sharma, P. Bharadwaj, and K.R. Seeja, "Privacy preserving data mining with 3-D rotation transformation," Journal of King Saud University-Computer and Information Sciences, Vol. 30, no. 4, pp. 524-530, 2018.

[15] T.I. Cannings, "Random projections: Data perturbation for classification problems," Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 13, no. 1, pp. e1499, 2021.

[16] S. Kotsuki, Y. Sato, and T. Miyoshi, "Data Assimilation for Climate Research: Model Parameter Estimation of Large-Scale Condensation Scheme," Journal of Geophysical Research: Atmospheres, Vol. 125, no. 1, pp. e2019JD031304, 2020.

[17] A. Rodriguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A.M. Mezher, J. Parra-Arnau, and J. Forne, "The Fast Maximum Distance to Average Vector (F-MDAV): An algorithm for k-anonymous microaggregation in big data," Engineering Applications of Artificial Intelligence, Vol. 90, pp. 103531, 2020.

[18] H. Zhou, G. Yang, Y. Xiang, Y. Bai, and W. Wang, "A Lightweight Matrix Factorization for Recommendation with Local Differential Privacy in Big Data," IEEE Transactions on Big Data, 2021.

[19] A. Dziedzic, & S. Krishnan, "Analysis of Random Perturbations for Robust Convolutional Neural Networks," arXiv preprint arXiv 2002.03080, 2020.

[20] J. Li, X. Kuang, S. Lin, X. Ma, & Y. Tang, "Privacy preservation for machine learning training and classification based on homomorphic encryption schemes," Information Sciences, Vol. 526, pp. 166-79, 2020.

[21] N.K. Anuar, A.A. Bakar, S. Yussof, F.A. Rahim, R. Ramli, & R. Ismail, "Privacy Preserving Features Selection for Data Mining using Machine Learning Algorithms," In2020 8th International Conference on Information Technology and Multimedia (ICIMU) IEEE, pp. 108-113, 2020.

[22] F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R.T. Leijenaar, A. Jochems, & P. Lambin, "Systematic review of privacy-preserving distributed machine learning from federated databases in health care," JCO clinical cancer informatics, Vol. 4, pp. 184-200, 2020.

[23] M. Cunha, R. Mendes, & J.P. Vilela, "A survey of privacy-preserving mechanisms for heterogeneous data types," Computer Science Review, Vol. 41, pp. 100403, 2021.

[24] G.N. Devi, & K. Manikandan, "Improved perturbation technique privacy-preserving rotation-based condensation algorithm for privacy preserving in big data stream using Internet of Things," Transactions on Emerging Telecommunications Technologies, Vol. 31, no. 12, 2020.

[25] N. Kousika, & K. Premalatha, "An improved privacy-preserving data mining technique using singular value decomposition with three-dimensional rotation data perturbation," The Journal of Supercomputing, Vol. 77, no. 9, pp. 10003-10011, 2021.

[26] G.S. Kumar & K. Premalatha, "Securing private information by data perturbation using statistical transformation with three dimensional shearing," Applied Soft Computing, Vol. 112, pp. 107819, 2021.

[27] M.A.P. Chamikara, P. Bertók, D. Liu, S. Camtepe, & I. Khalil, "Efficient privacy preservation of big data for accurate data mining," Information Sciences, Vol. 527, pp. 420-43, 2020.

[28] A. Kumar, R. Kumar, & S.S. Sodhi, "Intelligent privacy preservation electronic health record framework using soft computing," Journal of Information and Optimization Sciences, Vol. 41, no. 7, pp. 1615-1632, 2020.

[29] P. Bedi, and S.B. Goyal, "Privacy preserving on personalized medical data in cloud IoT using Extended Fully Homomorphic Encryption". 2022.

[30] V.S. Reddy, and B.T. Rao, "A combined clustering and geometric data perturbation approach for enriching privacy preservation of healthcare data in hybrid clouds," International Journal of Intelligent Engineering and Systems, Vol. 11, no. 1, pp. 201-210, 2018.

[31] S. Janakiraman, and D.P. Maruthakutty, "Advanced extreme learning machine-based ensemble classification scheme with enhanced data perturbation for human DNA sequences," Computational Intelligence, Vol. 37, no. 4, pp. 1890-1915, 2021.

[32] V. Santhana Marichamy, and V. Natarajan, "Efficient big data security analysis on HDFS based on combination of clustering and data perturbation algorithm using health care database," Journal of Intelligent & Fuzzy Systems Preprint, pp. 1-18.

[33] K. Sujatha, and V. Udayarani, "Chaotic geometric data perturbed and ensemble gradient homomorphic privacy preservation over big healthcare data," International Journal of System Assurance Engineering and Management, pp. 1-13, 2021.

[34] H. Chen, S. Das, J.M. Morgan, and K. Maharatna, "Prediction and classification of ventricular arrhythmia based on phase-space reconstruction and fuzzy c-means clustering," Computers in Biology and Medicine, Vol. 142, pp. 105180, 2022.

[35] A.K. Sahoo, S. Raj, C. Pradhan, B.S.P. Mishra, R.K. Barik, and A. Vidyarthi, "Perturbation-Based Fuzzified K-Mode Clustering Method for Privacy Preserving Recommender System," International Journal of Information Security and Privacy (IJISP), Vol. 16, no. 1, pp. 1-20, 2022.

[36] R.U. Haque, A.S.M. Hasan, T. Nishat, and M.A. Adnan, "Privacy-Preserving-Means Clustering over Blockchain-Based Encrypted IoMT Data," In Advances in Blockchain Technology for Cyber Physical Systems, 109-123, 2022. Springer, Cham.

[37] Z. Zhang, T. Wu, X. Sun, and J. Yu, "MPDP k-medoids: Multiple partition differential privacy preserving k-medoids clustering for data publishing in the Internet of Medical Things," International Journal of Distributed Sensor Networks, Vol. 17, no. 10, pp. 15501477211042543, 2021.

[38] M. Wang, W. Zhao, K. Cheng, Z. Wu, and J. Liu, "Homomorphic Encryption Based Privacy Preservation Scheme for DBSCAN Clustering," Electronics, Vol. 11, no. 7, pp. 1046, 2022.