E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 02-04-2025
Accepted: 13-05-2025
Published: 04-07-2025

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 3; Issue 3; 2025; Page No. 381-386

Secure Federated Learning for High-Dimensional Healthcare Data Using Differential Privacy

[1]Vishal Trivedi and [2]Dr. Sunil Bhutoda

[1]Research Scholar, P.K. University, Shivpuri, Madhya Pradesh, India
[2]Professor, P.K. University, Shivpuri, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.18413604

Corresponding Author: Vishal Trivedi

Abstract

Machine learning algorithms may be trained on decentralized data using Federated Learning (FL) when sharing raw data is not possible owing to privacy concerns. One example of this kind of data is EHRs, or electronic health records, which store private information about patients. Instead of sharing sensitive data, FL trains models locally and then aggregate their parameters on a central server. An effective method for training Machine Learning (ML) algorithms on distributed datasets when data owners are governed by restrictions that limit the sharing of raw data is Federated Learning (FL). There is less need to communicate raw data with people outside the premises with this strategy, which involves local training and model aggregation to a central server. Nevertheless, FL brings up valid issues around privacy. For that reason, we need more privacy safeguards. One state-of-the-art privacy technique is the differential privacy (DP) approach, which involves adding an extra layer of privacy by perturbing the local models before transmission. But this method could change the framework's usefulness. In order to strike a fair balance between privacy and usefulness, we employ a private method to clean raw data by combining DP noise with a top-down taxonomy tree. To train local models that may be shared in the FL architecture, the generalized data is utilized in conjunction with DP noise. The suggested architecture improves functionality while keeping the privacy budget low.

Keywords: Privacy Preservation, Federated Learning, Machine Learning, algorithms and architecture

1. Introduction

The latest developments in Generative AI are driven by Machine Learning (ML) methods, which are the backbone of descriptive, predictive, and prescriptive analysis. ML-based systems include recommendation systems, prediction engines, sentiment analysis services, object detection, anomaly or fraud detection, and others. The financial, pharmaceutical, and medical science industries aren't the only ones that can benefit. In fields like natural language processing and voice recognition, ML methods like Deep Neural Networks (DNN) have successfully tackled difficult problems. On the other hand, ML is used in the financial sector for things like client retention programs, algorithmic trading, and financial monitoring [11].

Generative AI, which depends on fundamental models-pre-trained models learned on massive quantities of data-and collaborative machine learning across domains are two examples of the more complex processes made possible by advances in pattern recognition and learning from single data sources. These advancements have completely changed the way business database insights are generated [12]. By ensuring that data is coming from several sources to cover diversity, volume, and dispersion, Machine Learning insights may be made exact, trustworthy, and efficient. When presented with unexpected data, machine learning models often underperform due to a lack of different training data sources and various perspectives. It must be noted that the data used to derive these conclusions is often limited to data owned by the company [14].

Consequently, the accuracy, efficacy, and consistency of these models may be compromised anytime they are used in different organizational settings. By embracing a free data exchange between many stakeholders, organizations may contribute to a varied dataset, increase the effectiveness of machine learning conclusions, and overcome these constraints. Improved accuracy and reliability in machine learning are outcomes of the expanded dataset's capacity to train models on a wider variety of data points. When data is shared, trends and patterns may be seen that would not be apparent when it is analyzed separately. By integrating data from several sources, organizations may uncover hidden insights and get a better understanding of intricate situations. When users share their data, it makes machine learning processes more open and accountable. Multiple stakeholders may access the created data and insights, which makes it simpler to check and evaluate the outcomes. The trustworthiness and dependability of machine learning insights are guaranteed by this openness, which in turn inspires confidence among stakeholders.

2. Literature Review

At present, the privacy preservation plays a vital role in data mining. Baozhen Lee *et al*. (2014) coordinated the process of protecting individuals' private while also handling their personal information, which necessitates a paradigm change in thinking about both publicity and privacy [9]. Along with measuring privacy using conventional standards, we provide a constraint on the nonlinear distortion's forecast accuracy. The main idea behind this method is that the user should be able to control the amount of privacy by adjusting the nonlinearity [15]. As an additional layer of defense against unauthorized access to critical database information, there is a tree-based approach known as a rapid perturbation algorithm [2].

According to Dinusha Vatsalan and Peter Christen (2016), privacy preservation approaches have several practical applications. suggested a privacy-preserving architecture that improves masking and matches patients with comparable medical histories. It finds the best values for data-dependent parameters and uses bloom filter encodings to conceal data. Using bloom filters, it conceals both numerical and string data. A large database with many attributes will increase the computational complexity of the suggested system [1]. Using these data, we may determine how comparable the attribute values are. Performance, data usefulness, and uncertainty or resistance to data mining algorithms are three ways to judge the success of privacy preservation algorithms. After learning about data breaches, users are understandably wary about disclosing any personally identifiable information.

A methodology was suggested by Samanthula *et al*. (2015) that conceals data access patterns, safeguards data confidentiality, and preserves the privacy of user input queries. That is, the semi-honest model is used to create the safe k-NN classifier over encrypted data. One method of protecting personal information is data concealing. The irretrievable PPDM issue is solved using the reversible privacy preserving data mining technique. The method of reversible data concealing is used [5]. To accomplish privacy preservation and knowledge verification, the privacy difference expansion (PDE) approach perturbs the original data and embeds it with a fragile watermark. It can also get the initial data back. Probabilistic information loss, privacy disclosure concerns, and classification accuracy are the metrics used to evaluate PDE performance [2].

Chen Yi Lin (2016) developed the RDT algorithm, which can both destroy and recover data. If you want to prevent data mining and knowledge reservation from revealing sensitive information, this algorithm is for you. To improve the adaptability of privacy-preserving measures, it makes use of a weighting system that can be adjusted and the level of data disruption [3]. To identify tampering with the disturbed data, a watermark might be included in the original data. Data loss and privacy breach are both mitigated by the suggested approach.

Sushmita Ruj & Amiya Nayak (2023), developed a distributed security architecture for smart grids that can aggregate data and restrict who may access it. Data aggregation safeguards consumers' personal information. Networks in the house, in buildings, and in nearby areas all work together to aggregate data. Cryptographic keys are distributed via a network of key distribution centers (KDCs) [6]. Use of attribute-based encryption (ABE) is key to the suggested access control method; this technology allows for limited access to consumer data held in data repositories and used by various smart grid users. The solution is resilient since the access control mechanism is distributed and does not depend on a single KDC to distribute the keys. Ensuring privacy while aggregating data and controlling access is the primary emphasis of this effort.

3. Materils and Methods

In this section, proposed methods detailing the privacy-preserving techniques will be discussed.

3.1 Problem Description

In this model (Fig. 1), we examine gene data of patients with diverse health disorders, such as heart failure, using the federated learning framework to discover possible risk factors. Due to health data regulatory rules, it is not possible to transfer this data across premises in its raw form, therefore data is gathered from several sites while keeping sufficient privacy. As a result, we construct the model locally and then upload it to the main server so it may be trained. The suggested method restricts communication with the central aggregated server to data pertaining to perturbed models [7]. The final model is built and trained by the trustworthy aggregator server using the aggregated local data. Privacy attacks, such as model inversion or reconstruction attacks, may still occur on model data communicated in a federated learning system [3, 4, 36].

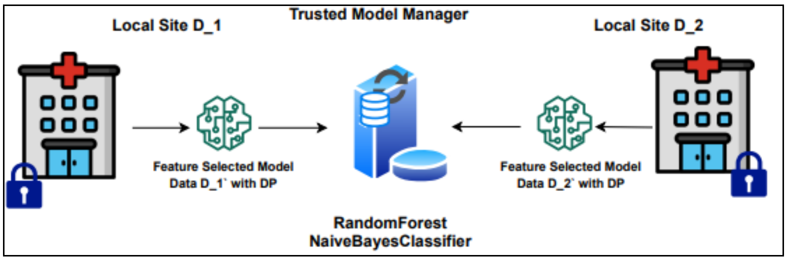


Fig 1: Multiple Data Owners are training a model collaboratively using federated machine learning algorithm providing a privacy guarantee over the data

Thus, it is necessary to provide an extra degree of privacy. By introducing noise into the model data throughout the sharing process between sites and aggregator servers, the differential privacy algorithm accomplishes the privacy mechanism [8].

If there is a significant quantity of noise introduced to high-dimensional data with more columns, the accuracy of the model may be affected. In order to increase the accuracy of the model and minimize the amount of data dimensions, a federated correlation-based feature selection is devised to obtain a shared list of the features that are used for training.

One reason people have heart failure is due to transthyretin amyloid cardiomyopathy. An ML model is a reliable tool for predicting the likelihood of certain health problems. Thus, in order to forecast the likelihood of heart failure, our federated machine learning model will examine the patient's genetic information and determine if a cohort of individuals may be at risk of wild-type transthyretin amyloid cardiomyopathy, based on established characteristics. The dataset is divided into two parts, each including 1713 samples. In the first part, 855 samples are classified as ATTR-CM, or wild-type amyloidogenic TTR cardiomyopathy. In the second portion, there are 858 samples, with 1874 phenotypes (features) per sample. Our produced model should be able to use this dataset to foretell if the patient will develop heart failure using privacy-preserving federated learning [9].

Due to their massive size, electronic health records and databases may not be practical when dealing with constrained privacy resources. To make the most of the limited privacy budget, we reduce the dimensionality by using feature selection to eliminate unnecessary columns. Finding the columns in the given data with the strongest link to the condition is the first step in our feature selection process. We use the Laplace Transformation to introduce noise into the resulting statistics and Differential Privacy (DP) methods to guarantee privacy. The resultant noisy data is sent to the aggregator server instead of the raw data. The data is then used by the central aggregator server to educate a machine learning framework, which acquires a worldwide model for predicting the probability of heart failure [1].

3.2 Feature Selection

Many genes available in a genomic dataset are the first essential factor impacting the model's utility due to diversity in the whole dataset. Any machine learning system processing such vast data dimension often leads to less accuracy because some genes provide no value to the analysis. There- fore, proper methodologies for feature selection from the whole dataset constantly improve the score efficiently.

While reducing the data dimension, we focus on two things: Choosing genes based on their correlation with the disease and sending only summarized data as a model to the central Server. A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables [2].

$$corr(X,Y) = \rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu x)(Y-\mu y)]}{\sigma_X \sigma_Y}$$

Similarly, in our approach, each local server chooses a list of a maximum of 200-300 Genes that have the highest correlation with the disease as per the above equation to build up the local model. Finally, the central Server receives and matches each local model's columns and feeds them to the ML framework for training.

In our approach, reduced dimensions require less privacy budget (in terms of ε) while maintaining the model's utility. Within a limited budget, if the number of columns is high, the privacy budget will be distributed to each column with less amount, resulting in more noise to be added. Therefore, more noise can impact the accuracy of the score directly. After feature selection in our approach, the data dimension reduces, resulting in higher accuracy through less noise added [4].

According to Table 1, if any gene from the whole data set contains control value as *True*, refers to *Positive Prediction of Heart Failure* while *False* value represents negative prediction. We calculate the correlation value for each column and define it as (μ, σ) for *True* prediction label and (μ', σ') as the *False* prediction label. Then we select the top 200-300 columns based on the correlation value [5]. In this way, we calculate the feature for each column and generate the two rows of data based on the control value as per shown in Table 1. Note that these two rows of data do not contain the actual raw value, rather the calculated model data based on the highest correlation for that column. Finally, after noise addition, Table 2 is sent to the model manager or central aggregation server with additional layer of protection using differential privacy.

3.3 Privacy Mechanism

Our approach solely depends on the differential privacy mechanism to maintain the additional privacy layer of the shared model data. In this approach, the noise will be added to the model data based on *Laplace Mechanism*. The privacy budget, ε, is varied based on the data dimension to achieve the best utility.

Table 1: Column Selection based on highest correlation value

***Gene*$_1$**	***Gene*$_2$**	***Gene*$_3$**	**Control Value**
(μ_1, σ_1)	(μ_2, σ_2)	(μ_3, σ_3)	# Have Heart Disease
($\mu'1$, $\sigma1'$)	($\mu'2$, $\sigma2'$)	($\mu3'$, $\sigma3'$)	# Don't have heart disease

Table 2: Differential Privacy mechanism applied to the model data

***Gene*$_1$**	***Gene*$_2$**	***Gene*$_3$**	**Control Value**
(μ_1, σ_1) + $Lap_{\Delta}f/\varepsilon_1$	(μ_2, σ_2) + $Lap_{\Delta}f/\varepsilon_2$	(μ_3, σ_3) + $Lap_{\Delta}f/\varepsilon_3$	Have Heart Disease
($\mu1'$, $\sigma1'$) + $Lap_{\Delta}f/\varepsilon_1$	($\mu'2$, $\sigma2'$) + $Lap_{\Delta}f/\varepsilon_2$	($\mu3'$, $\sigma3'$) + $Lap_{\Delta}f/\varepsilon_3$	Don't have heart disease

Data size will be reduced by dimension first based on feature selection rather than applying noise to the complete data sets. Finally, the noise will be added to the summarized data, as shown in Table 2. Note that the total privacy budget ε is distributed to each column based on the inverse correlation value. Therefore, the column with the highest correlation value will receive the most privacy budget and have less noise to be added [7]. Also, according to the composition theorem of the DP mechanism as detailed in the Background section, If $F_1(x)$ satisfies ε_1-differential privacy and $F_2(x)$ satisfies ε_2-differential privacy, then the mechanism $G(x) = (F_1(x), F_2(x))$ which releases both results satisfies $\varepsilon_1 + \varepsilon_2$-differential privacy satisfying the following equation:

$$\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n = \varepsilon,\ \varepsilon_i \propto (Corr\ (\mu_i, \sigma_i))^{-1}$$

Therefore, the total privacy budget will be equal to the summation of distributed budget for each column. In this way, the more relevant column with the disease prediction will add less noise value, resulting in an improved utility of the framework.

3.4 Federated Learning Mechanism

In our federated framework, after adding the privacy mechanism, only model data from each local data owner are sent to the aggregator server to train the ML framework. We have utilized two ML algorithms for the federated training: a) Naive Bayes Classifier and b) Random Forest in a federated setting to validate the efficacy of the proposed method [9].

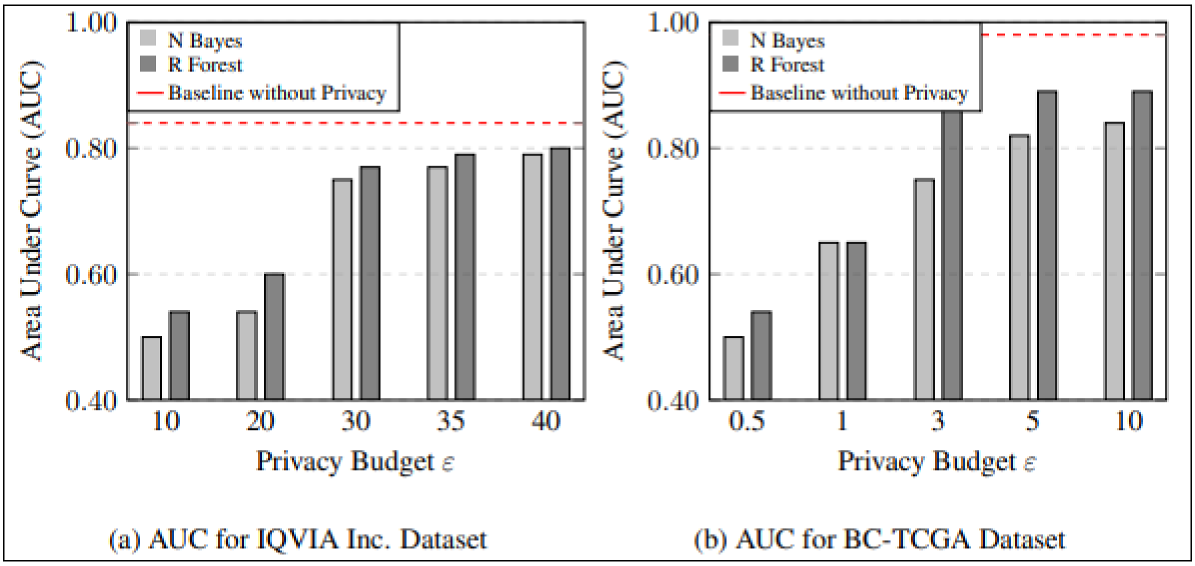
The data owners remain solely responsible for building their local model, while the aggregator server builds the results in a collaborative learning setting. The statistics required for Random Forest and Naive Bayes, for example, are first completed at the data owner's location and then shared with an additional privacy layer (DP) to be robust against further model inversion attacks. For Random Forest, after reducing the column with the highest correlation values, only the two rows with control value *True* or *False* is calculated as (μ, σ) and (μ', σ') respectively. Finally, Table 2 is generated at each data owner and sent to the aggregator server after adding the noise according to the Laplace mechanism. The same approach is applied for Random Forest, except the model is built on the Tree from raw data and noise is added afterwards.

4. Results

In this section, the experimental result is described. Since the proposed method is a generalized data-sharing mechanism for federated ML applications, we experiment with different settings as portrayed in Table 3. We utilized multiple machines at our lab as server-client settings to conduct the experiments. The average latency between the servers was minimal [12].

4.1 Experimental Setup

The experimental data were taken from the iDASH 2021 competition which tested the proposed solutions with a single dataset: IQVIA Inc, for predicting causes of certain heart failure. We utilized additional datasets from BC-TCGA for cancer prediction alongside to compare our proposed method:



(a) AUC for IQVIA Inc. Dataset

(b) AUC for BC-TCGA Dataset

Fig 2: Accuracy difference with different privacy budgets and methods

Table 3: Different experimental parameters considered in this approach

Dataset	ML Methods	budget ε	Dimensions
IQVIA Inc.	Naive Bayes	[20, 30, 40, 50, 60]	1874
BC-TCGA	Random Forest	[1, 3, 5, 10, 20]	17814

IQVIA Inc.: 1713 samples, where 855 samples are diagnosed as wild-type amyloidogenic TTR cardiomyopathy (ATTR-CM) as well as positive cases of heart failure, and 876 negative controls.

BC-TCGA: 17814 genes with 424 positive labels and 50 negative labels.

The training data for the ML models were chosen at random in an 80:20 split, with 80% of the data being used in training. In a two-party setup where the data is split into two, the identical training data was used in both Naive Bayes and Random Forest [14]. The training procedure was repeated ten times, with the Area Under the Curve (AUC) values from each test set being averaged. We also experimented with different parameters, privacy budget ϵ varying data dimension to 200-350 for IQVIA and 20-50 for BC-TCGA. These are outlined in Table 3.

4.2 Accuracy

To measure the two ML models' significant accuracy, we used the Area Under Curve (AUC) statistic. Since the curve (receiver operating characteristic) is generated by True Positive and False Positive rates, AUC appropriately characterizes the binary classifier. Moreover, it selects many thresholds between 0 and 1, with 1 being the most accurate; this means that the model accurately predicts the presence of all data points for all thresholds. The model is no more accurate than a coin flip for any binary classification when the area under the curve (AUC) is less than or equal to half [15]. Here, the suggested model is limited to correctly classifying positive data and fails miserably when faced with negative ones. Our relationship between privacy budget (ε) and Area Under Curve (AUC) for two separate approaches, Naive Bayes and Random Forest, is shown in Figure 2. Results demonstrate that using a privacy budget of 30 or more yields better AUC values when applying the Random Forest algorithm on the competition's (IQVIA Inc.) dataset. The experiment is designed with the reduced dimension set at m′ = 250. Figure 2 shows a similar trend; with privacy budget 10, we were able to analyze BC-TCGA datasets with more accuracy. In this scenario, the reduced dimension is defined as m′ = 20. Lastly, when it comes to accuracy, we see a similar trend, where higher AUCs are produced by larger ε values [6].

Similarly, when looking at AUC, Fig. 2 reveals that Random Forest consistently produces higher AUCs regardless of the value of m′. We found that the Random Forest method takes more time to run than Naive Bayes due to the increased number of calculations it required, in addition to the AUC discrepancy. The findings are negatively affected by both high and small values of m′, which shows how the data behaves with several dimensions. Lower values of m′ lead to significant data loss, while higher values of m′ take up a lot of ε and introduce too much noise into the data. Hence, for ε = 30, m′ ∈ {250, 300, 350} works effectively.

Using BC-TCGA data, we implemented our solution. It should be noted that the total score seems to be close to 98% for the BC-TCGA data, whereas the IQVIA data may only approach 80% with the DP option. Reason being, in the same configuration, BC-TCGA data achieves a maximum baseline score of 98%, but IQVIA data achieves a maximum baseline score of 84% when no privacy protection is included in the central architecture. In addition to more precise removal of features, IQVIA data consists entirely of binary values, which is called Haplotype Data [5]. So, in order to attain the same data usefulness as BC-TCGA, this data set with reduced dimensions influences the total score and requires a larger privacy budget. We hypothesise that the accuracy of the framework was compromised as a result of data loss in the haplotype gene data, which led to this situation. Because more relevant data were selected using feature selection to reduce dimension size, BC-TCGA data had no impact on the score.

5. Conclusion

In this study, we present three approaches for conducting a secure analysis of sensitive health- care data that is distributed among various participants. These frameworks are designed to perform ML classification in both horizontally and vertically partitioned data with an adequate level of utility while maintaining privacy. We proposed a distributed machine learning framework while guaranteeing privacy on vertically segregated data. In our method, each client uses LR and LSTM neural networks to make local predictions based solely on local features. Then, to offer an extra layer of privacy, a certain amount of noise is added to the prediction results using the DP algorithm. In addition, the weighted feature function, which is computed based on local feature sets, is applied to the final prediction. The central server then receives the perturbed scores along with a proper weight to calculate the final prediction. No raw data, features or model parameters are shared in any phase of the training. The results show that the federated version of the algorithm that uses encrypted gradients performs almost as well as its unencrypted counterpart, thus proving that partially homomorphic encryption is a viable tool that can be used to implement privacy in matrix decomposition based collaborative filtering methods without compromising much on accuracy as compared to its version that communicates gradients using plaintext to the server.

6. References

1. Han S, Ng WK, Wan L, Lee V. Privacy-preserving gradient-descent methods. IEEE Transactions on Knowledge and Data Engineering. 2020;22(6):884–899.
2. Vatsalan D, Christen P. Privacy-preserving matching of similar patients. Journal of Biomedical Informatics. 2016;59:285–298.
3. Couellan N, Jan S, Jorquera T, George JP. Self-adaptive support vector machine: A multi-agent optimization perspective. Expert Systems with Applications. 2015;42(9):4284–4298.
4. Ruj S, Nayak A. A decentralized security framework for data aggregation and access control in smart grids. IEEE Transactions on Smart Grid. 2023;4(1):196–205.
5. Nguyen T, Khosravi A, Creighton D, Nahavandi S. Medical data classification using interval type-2 fuzzy logic system and wavelets. Applied Soft Computing. 2015;30:812–822.
6. Zhu T, Ren Y, Zhou W, Rong J, Xiong P. An effective privacy-preserving algorithm for neighborhood-based collaborative filtering. Future Generation Computer Systems. 2014;36:142–155.
7. Wang T, Huang H, Tian S, Xu J. Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels. Expert Systems with Applications. 2020;37(9):6663–6668.
8. Chen TS, Lee WB, Chen J, Kao YH, Hou PW. Reversible privacy-preserving data mining: A combination of difference expansion and privacy preserving. The Journal of Supercomputing. 2023;66(2):907–917.
9. Vaidya J, Shafiq B, Fan W, Mehmood D, Lorenzi D. A random decision tree framework for privacy-preserving data mining. IEEE Transactions on Dependable and Secure Computing. 2014;11(5):399–411.
10. George VS, Raj VC. Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. International Journal of Computational Science Engineering Survey. 2021;2(3):16–27.
11. Vinodhini G, Chandrasekaran RM. A comparative

performance evaluation of neural network-based approaches for sentiment classification of online reviews. Journal of King Saud University – Computer and Information Sciences. 2016;28(1):2–12.

12. Nedic V, Cvetanovic S, Despotovic D, Despotovic M. Data mining with various optimization methods. Expert Systems with Applications. 2014;41(8):3993–3999.
13. Wang S, Li Z, Liu C, Zhang X, Zhang H. Training data reduction to speed up SVM training. Applied Intelligence. 2014;41(2):405–420.
14. Xiao X, Tao Y. Personalized privacy preservation. In: Proceedings of the ACM SIGMOD International Conference on Management of Data; c2016.
15. Xiao X, Tao Y, Chen M. Optimal random perturbation at multiple privacy levels. Proceedings of the VLDB Endowment. 2019;2(1):814–825.