Briefings in Bioinformatics, 2022, **23**, 1–12

https://doi.org/10.1093/bib/bbab498

Problem Solving Protocol

FedSPL: federated self-paced learning for privacy-preserving disease diagnosis

Qingyong Wang and Yun Zhou

Corresponding author: Yun Zhou, Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China; E-mail: zhouyun@nudt.edu.cn

Abstract

The growing expansion of data availability in medical fields could help improve the performance of machine learning methods. However, with healthcare data, using multi-institutional datasets is challenging due to privacy and security concerns. Therefore, privacy-preserving machine learning methods are required. Thus, we use a federated learning model to train a shared global model, which is a central server that does not contain private data, and all clients maintain the sensitive data in their own institutions. The scattered training data are connected to improve model performance, while preserving data privacy. However, in the federated training procedure, data errors or noise can reduce learning performance. Therefore, we introduce the self-paced learning, which can effectively select high-confidence samples and drop high noisy samples to improve the performances of the training model and reduce the risk of data privacy leakage. We propose the federated self-paced learning (FedSPL), which combines the advantage of federated learning and self-paced learning. The proposed FedSPL model was evaluated on gene expression data distributed across different institutions where the privacy concerns must be considered. The results demonstrate that the proposed FedSPL model is secure, i.e. it does not expose the original record to other parties, and the computational overhead during training is acceptable. Compared with learning methods based on the local data of all parties, the proposed model can significantly improve the predicted F1-score by approximately 4.3%. We believe that the proposed method has the potential to benefit clinicians in gene selections and disease prognosis.

Keywords: disease classification, data privacy, federated learning, self-paced learning, gene selection

Introduction

Recently, the high incidence of cancer cases is a significant social and medical concern. Early disease detection is an effective way to minimize the fatality rate, and machine learning is typically employed to facilitate detection and find the relationships among instance features. Various biomedical and machine learning algorithms have been used in disease prediagnosis, e.g. the hierarchical fully convolutional network [1] and disease prognosis prediction [2]. In addition, a variety of early disease detection models have been proposed, e.g. the structured sparsity regularization model [3], and semi-supervised high-quality instance selection in pseudo-labels [4].

Identifying key genes is important in early disease detection, and various machine learning algorithms, including semi-supervised labeled learning and self-paced learning, have proposed and applied gene feature variables and diagnosis information from patient history data [5]. These methods are broadly supervised learning styles and require many samples. However, most cancer samples only include a few tumor samples because labeled samples are very costly, which can introduce overfitting in learning algorithms [6].

Various methods have been proposed to solve this limited samples problem, including data augmentation [7], data fusion [8] and the ensemble classifier [9]. Such methods can improve model performance efficiently; however, integrating data directly may result in privacy leakage because the data are distributed in different areas. Therefore, preserving privacy is a key challenge to train these samples' distribution of diverse clients.

Privacy-preserving methods have been proposed to solve this problem, including multikey privacy preservation via deep learning [10], privacy-preserving k-means clustering [11, 12] and data security adoption cloud computing [13]. Generally, differential privacy [14] does not influence the outcome of any final analysis even when a single noisy record is removed or added. Such methods consider unique key information protection. However, we hope that training hyperparameters could be preserved.

Thus, federated learning was introduced. Federated learning is a method to maintain all associated privacy data in local institutions where the data belong and only deliver hyperparameters in a shared global model. In

Qingyong Wang is a Ph.D. student with Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, China. His research interests are machine learning and bioinformatics.

Yun Zhou received the Ph.D. degree in Computer Science from the Queen Mary University of London, in 2015. He is currently an Associate Professor with the Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, China. His research interests are data mining and probabilistic graphical models. He applies these techniques to a wide range of real-world problems. He has published several papers in reputed journals and conferences in this area, including ICDM, IJCAI, IJAR, UAI, and PGM.

Received: July 29, 2021. **Revised:** October 28, 2021. **Accepted:** October 30, 2021

addition, federated learning can construct a global model based on datasets distributed across multiple devices without exposing the raw data [15]. Federated learning has been used in numerous applications, e.g. drug discovery with quantitative structure–activity relationship prototype analysis [16], privacy-preserving disease classification with blood transcriptomes and X-ray images [17] and federated semi-supervised learning of COVID chest computed tomography [18]. A federated learning privacy-preserving framework using a decentralized programming paradigm and secure multiparty computation was proposed for a healthcare system [19]. In addition, a federated deep learning algorithm for heart activity data was developed very recently to preserve the privacy of health data collected from individuals [20].

Noisy data are common in many real-world applications. If training these original data directly, which may result in overfitted models of federated learning. Recently, the self-paced learning (SPL) approach, which mimicks the cognitive mechanism of children to gradually learn from simple to difficult targets, was initially proposed [21] to retain high-quality samples and remove noisy samples. The SPL approach can prevent the influence of noisy samples on model performance and been applied in disease diagnosis [22], drug–target interactions prediction [23] and molecular descriptor selection [24]. In addition, it can only uses a part of the crucial data to improve model performances and reduce the risk of privacy disclosure. Thus, this learning technique is suitable in the healthcare work.

In this paper, we exploit the virtues of federated learning and SPL to improve model performance under privacy-preserving data sharing to resist data noise of federated learning in disease diagnosis. Specifically, we first employ SPL to initially select high-confidence datasets (thereby avoiding noisy samples) and reduce data exposure problems to the maximum extent. Federated learning is used to learn various data distributed across different clients. Experiments were conducted on gene expression data, and the results demonstrate that the proposed algorithmic framework can improve performance under privacy protection than existing learning methods, e.g. federated averaging (FedAvg). The embedded classifier with a federated learning method under privacy protection can train a model for disease classification and gene selection more effectively [25]. Our primary contributions are summarized as follow.

- We proposed a decentralized and collaborative privacy-preserving framework that allows doctors to exploit the benefits of sharing rich private healthcare data while conserving patient privacy.
- We presented the privacy-preserving federated self-paced learning (FedSPL) method to effectively select samples from high to low confidence in each individual client. The proposed FedSPL can avoid degrade performance of the global model from noisy data.
- We present the results of extensive experiments conducted to evaluate the effectiveness and the significance of the proposed method. The results indicate that the effectiveness of the proposed method is higher than that of the compared methods, and the proposed model can select small and highly relevant genes to facilitate early disease prognosis.

The remainder of this paper is organized as follows. Section 2 introduces related work. In Section 3, we introduce the FedSPL model to protect privacy and iterate noisy samples to improve the model performance. Section 4 presents experimental results involved a pure data characteristic and experiments conducted on four gene expression datasets for verification. Conclusions and suggestions for future work are presented in Section 5.

Related Work

Various machine learning methods are used in prediagnosis and prognosis. For example, a deep learning-based framework with denoising autoencoder has been proposed to predict cancer prognosis accurately using multi-omics data analysis [26]. The evolution algorithm with multiple classifiers has also been proposed to remedy the class-imbalance problem in medical and biological datasets [27].

Privacy-preserving data mining has been studied extensively for decades and has been applied to various tasks, e.g. such as privacy-preserving mining of association rules [28], clustering [29] and deep learning [30]. Differential privacy has been widely used to privacy preserving in many different domains. For example, deep learning with differential privacy was developed to protect sensitive information [31]. In addition, a practical privacy-preserving deep learning system that does not share raw data was proposed to learn an accurate neural network model jointly multiple parties [32]. Federated learning has received increasing attention for privacy protection because it can enable participants to learn a global shared model collaboratively without revealing their data. Since then, many efforts have been devoted to promote federated learning [33].

Federated learning has recently received significant attention in the healthcare domain. For example, predictive model based on federated learning through peer-to-peer collaboration without raw data exchange was proposed to accurately predict heart-related hospitalizations [34]. In the drug discovery field, federated learning has been used to establish a joint global model with different pharmaceutical companies without sharing local data to predict molecular properties [35]. In particular, federated learning methods can accelerate COVID-19 drug repurposing because it can train models across decentralized edge devices or servers hosting different local samples [36]. In addition, an

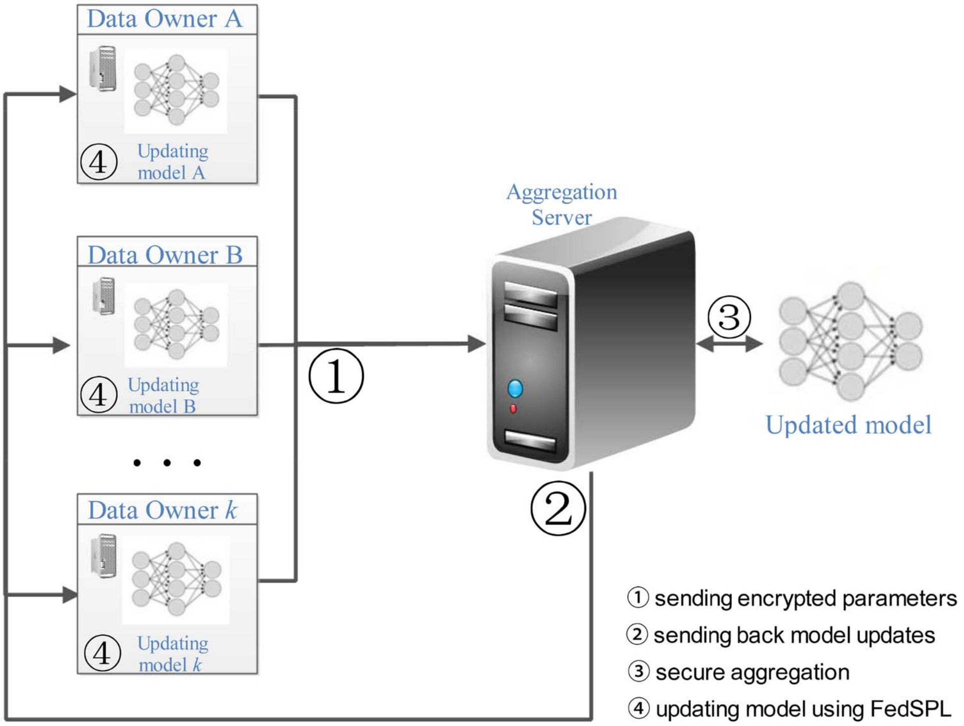


Figure 1. Workflow of FedSPL. An individual client trains the personal model parameters on its own private data, and encrypted model parameters are sent to the server. Then, the server performs secure aggregation and send the aggregated results back to each client. Finally, each client updates its own model using the decrypted. parameters.

efficient privacy-enhanced federated learning scheme was proposed previously. This scheme can effectively prevent private data leakage even if multiple participants collaborate [37]. Existed problems related to isolated data islands and models trained on the cloud fail on personalization in smart healthcare and such problems are addressed by the federated transfer framework (FedHealth) to perform data aggregation and build relatively personalized models using federated learning and transfer learning, respectively [38]. Although federated learning has outstanding advantages, it also has some problems, e.g. training with noisy data, which must be addressed. For example, Tolpegin *et al.* [39] proposed a against attacks strategy to identify malicious participants of federated learning to circumvent data noises.

FedAvg (i.e. Federated Averaging), which is typical federated learning method that is used in various fields, was proposed to train an aggregate model without uploading client data to a server using averaged parameters on the server [40]. In addition, McMahan *et al.* [41] presented a FedAvg framework to detect COVID-19 chest X-ray images using deep convolutional neural networks, which allows clinicians all over the world to reap the benefits of sharing rich private medical data while conserving privacy. The recently released information about privacy-preserving machine learning attempts to accomplish accelerated development of a flexible and secure federated learning framework. There are many open-source codes or frameworks for federated learning for product and research development. For example, PySyft, which is a Python library that uses federated learning in PyTorch or TensorFlow platform, provides secure and private deep learning with differential privacy, encrypted computation and homomorphic encryption [42]. Federated AI Technology Enabler is another secured computing architecture that supports numerous federated learning architectures and machine learning algorithms [43].

SPL is a reweight learning strategy that has been used in many fields in recent years, including healthcare and clinical medicine recognition [44]. For example, a deep active SPL model has been proposed to reduce annotation effort in biomedical images [45]. Self-paced balance learning was proposed to recognize class-imbalanced clinical skin disease data [46]. In addition, Xia *et al.* [23] proposed a collaborative matrix method to predict drug–target interactions.

As mentioned previous, noise in data is common real-world applications; thus, researchers have also used feature selection methods to select key subset features and reduce the influence of redundant variables that might contain noises. Currently, there are four main methods, i.e. filter methods [47], wrapper methods [48], embedded methods [49] and combined methods [50, 51]. Filter methods typically function as a preprocessing step prior to model training. In contrast, wrapper approaches evaluate the relevant features using a prediction method. Previous studies have shown that the embedded method can achieve excellent performance because it can combine the benefits of filters and wrappers [52].

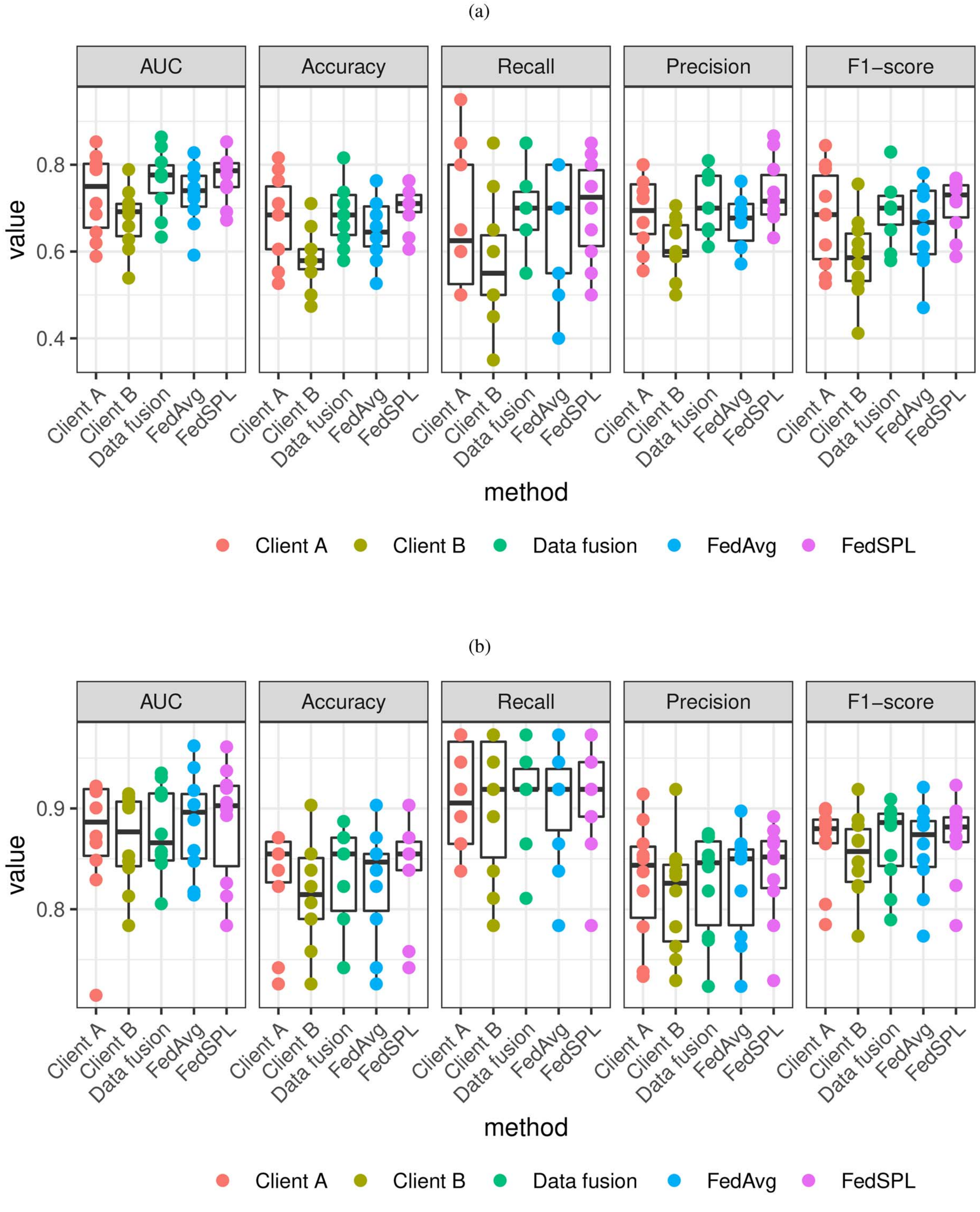


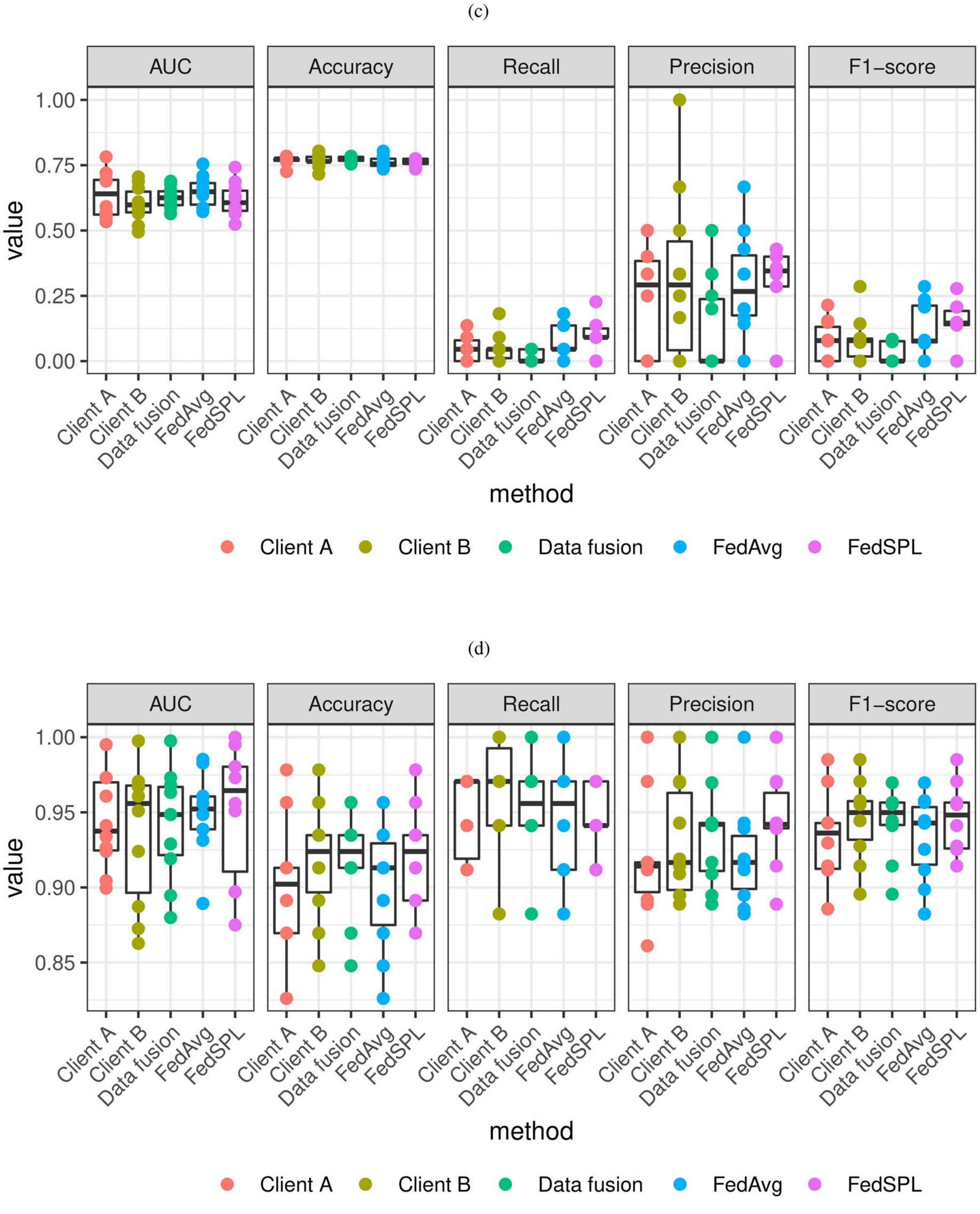
Figure 2. Boxplots showing the performance (AUC, Accuracy, Recall, Precision and F1-score) of five learning methods obtained on different the (**a**) GSE4115, (**b**) GSE21050, (**c**) GSE25066, (**d**) GSE30784 datasets. The horizontal axis shows the specific algorithm, and the vertical axis shows the performance of algorithm for different metrics. Each boxplot is a graphical summary of a distribution of the best learned values by multiple runs for each algorithm. The length of the boxplot represents the variance. When the length is shorter and compact this indicates that the deviation between results is small. Therefore, the best boxplot among all algorithms is the most compact and higher. Generally, we found that the proposed method outperformed the compared methods on four datasets.

Methods

In real-world federated machine learning applications, we typically assume that most participating clients are not malicious. The important characteristic of federated learning is training a global model without revealing the input data or the output of the model to other clients to secure private information.

Federated self-paced learning

Federated learning involves learning a global model via a coordinated central server while the raw data are stored at the local clients, who only send the clients' model parameters. Here, let G be a global model and let $L = \{l_k\}_{k=1}^{K}$ be a set of local models for K clients. In addition, let $D = \{\mathbf{x}_j, y_j\}_{j=1}^{m}$ be a given dataset, where $\mathbf{x}_j$ is an


(c)

(d)

Figure 2. Continued.

arbitrary training example with corresponding one-hot label $y_j \in \{1, 2, ..., C\}$ for the C-way multi-class classification problem and m is the number of examples. D comprises K sub-datasets $D^{l_k} = \{\mathbf{x}_j^{l_k}, y_j^{l_k}\}_{j=1}^{m^{l_k}}$ privately collected at each client or local model l_k. In each communication round t, G first randomly selects K local models that are available to train $L^t \subset L$ and $|L^t| = K$. Then, the global model G then initializes L^t with global weights w^G, and the active local models $l_k \in L^t$ perform supervised learning to minimize loss $L(w^{l_k})$ on the corresponding subset D^{l_k}. G then aggregates the learned weights $w^G \leftarrow \frac{m^{l_k}}{m} \sum_k^K w^{l_k}$ and broadcasts the weights of the new aggregation to the local models that would be available at the next round $t + 1$. This learning strategy is repeated until the final round R. Here, the goal is to solve Equation (1).

$$argmin_{w \in R^d} \left[l_w \triangleq \frac{1}{K} \sum_{k \in [K]} L_k(w) \right] \tag{1}$$

Given each individual training dataset $D = (\mathbf{x}_j^{l_k}, y_j^{l_k})_{j=1}^{m^{l_k}}$ with m^{l_k} samples, where $\mathbf{x}_j^{l_k} \in R^d$ represents the *j-th* sample, $y_j^{l_k}$is its corresponding output (e.g. $y_j^{l_k} \in \{0, 1\}$ for binary classification problem). Here, let $f(\mathbf{x}_j^{l_k}, w^{l_k})$ represents the learned function and the model parameter set w^{l_k}. In addition, $L(y_j^{l_k}, f(\mathbf{x}_j^{l_k}, w^{l_k}))$ is the loss function of the *j-th* sample. The optional objective of SPL is to collectively optimize the model parameter w^{l_k} and latent sample weights $v = [v_1, v_2, ..., v_m^{l_k}]$ via the following minimization problem:

$$\min_{w^{l_k}, v} E(w^{l_k}, v; \lambda) = \sum_{j=1}^{m^{l_k}} v_j L(y_i, f(\mathbf{x}_j, w^{l_k})) + g(\lambda, v_j), \qquad (2)$$

where $g(\lambda, v_j)$ represents self-paced regularize and λ is a penalty that controls the learning pace.

Additionally, we integrated the advantages of the SPL method and federated learning techniques to address the above issues. The procedure of the our proposed approach is illustrated in Algorithm 1, the condition of model converge is when the achieved the largest performances in training datasets. In particular, high-confidences samples $\{(\hat{\mathbf{x}}^{l_k}, \hat{y}^{l_k})\}$ are selected first by SPL to improve performance and protect privacy. Then, the embedded approach is employed to train datasets for gene selection.

The FedSPL system is shown in Figure 1. As can be seen, all participants uses the same data structure to learn a machine learning pattern collaboratively with the help of a centralized server.

Optimization algorithm

Here, we present an alternating optimization algorithm to solve Equation (2). This algorithm optimizes each component iteratively while keeping the others fixed. Given sample weights v, the kernel problem of the minimization over w is to decrease the value of the weighted loss function, independent on regularizer $g(\lambda, v)$. For model parameter w, the optimal weight of the *j-th* sample is determined in Equation (3),

$$min_{v_j} v_j L(y_j, f(\mathbf{x}_j, w)) + g(\lambda, v_j) \qquad (3)$$

Since $l_j = L(y_j, f(\mathbf{x}_j, w))$ indicates constant once w is given, the optimal value of v_j indicates uniquely determined by the corresponding minimizer function $\sigma(\lambda, l_j)$ that satisfies the following.

$$\sigma(\lambda, l_j) + g(\lambda, \sigma(\lambda, l_j)) \le v_j l_j + g(\lambda, v_j), \forall v_j \in [0, 1] \qquad (4)$$

For instance, if $g(\lambda, v_j) = -\lambda v_j$ [21], the optimal v_j^* is determined by

$$v_j^* = \sigma(\lambda, l_j) = \begin{cases} 1 & if \quad l_j \le \lambda \\ 0 & otherwise \end{cases} \qquad (5)$$

By increasing the value of λ and μ progressively so that more samples will be included in the next iteration, increasingly low-quality instances are contained in the training model procedure until the latest iteration of the model obtain the best performance. Existing multifarious researches have been contributed to learn appropriate minimizer self-paced regularizers [53, 54], according to require the explicit form of regularizer $g(\lambda, v)$. $\sigma(\lambda, l)$ is then derived from the form of $g(\lambda, v)$. Finally, the newest samples with high confidence were used in the following.

Experimental Results

In this section, we further evaluated the proposed method by implementing a FedSPL prototype on four benchmark datasets, an internet public variable datasets from US National Library of Medicine National Institutes of Health; furthermore, details are described in Table 1. The experiments were conducted to analyze the learning processes and performance improvement.

Table 1 shows the details of training dataset, which was divided into training data and testing data. In this work, the number of clients K was set two. For example, the lung cancer dataset (Dataset ID: GSE4115) includes 187 samples, among individual 75 samples distributed in Clients A and B, respectively. The testing dataset included 37 samples and 22 228 genes. The compared methods include client A, client B, FedAvg and data fusion (i.e. local learning where data are combined directly without considering privacy preserving). In model training, as slowly increase of the training model parameter λ, the accurate training rate is steadily changed. The latest iterative samples with excellent performance were selected to learn the following steps in the FedSPL. In order to assess the classification performances, five metrics: accuracy, AUC, recall, precision, F1-score are used here.

The accuracy, AUC, recall, precision and F1-score were used to compare in client A, client B, data fusion, FedAvg and the proposed FedSPL method. Table 2 by the line of comparison describes shows that F1-score of FedSPL approach is greater than that of the compared approaches, not only is individual clients' result but also are results of data fusion, FedAvg. The comparison results demonstrate that the proposed method not only can protect privacy but also can improve performance efficiently. The best F1-score value obtained by the FedSPL model was 95.0% on the GSE30784 dataset, which is approximately 2% greater than the worst F1-score value, which was obtained by FedAvg (93.3%). The highest improved F1-score value obtained by FedSPL was 19.9% on the GSE25066 dataset. The previous value was 7.9%

Algorithm 1: FedSPL. The K clients are indexed by k. B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

```
/* Server */
initialize $w_0$;
for each round $t$ =1,2,...,$R$ do
    $m \leftarrow max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of K $clients$)
    for each client $k \in S_t$ in parallel do
        $w^{l_k}_{t+1} \leftarrow$ ClientUpdate $(k, w^{l_k}_t)$
        $\bar{w}_{t+1} \leftarrow \sum_{k=1}^{K} \frac{m^{l_k}}{m} w^{l_k}_{t+1}$
    end
end
---------------------------------------------------------------
/* Client */
ClientUpdate$(k, w^{l_k})$ / / Run on client $k$
$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
for each local epoch $i$ from 1 to $E$ do
    Updated B by $\lambda$: Obtain optimal pseudo label for each of selected examples by solving
      $\lambda^* = argmin_\lambda \sum v^*_j L(y^{l_k}_j, g(\mathbf{x}^{l_k}_j; w^{l_k})) + g(\lambda, v^*_j)$ ;
    if $\lambda^*$ is small then
        increase with $\lambda$ by the stepsize $\mu$;
    end
    $\lambda$ is obtained to solve Equation (2), next updated training samples are obtained to train a model.
    for batch $b$ in $B$ do
        $w^{l_k}_{b+1} \leftarrow w^{l_k}_b - \eta \nabla l(w^{l_k}; b)$
    end
end
Return $w^{l_k}$ to server
```

Table 1. Summary of datasets (GSE4115, GSE21050, GSE25066 and GSE30784)

Dataset ID	Samples	Client A	Client B	Test samples	Genes
GSE4115	187	75	75	37	22228
GSE21050	310	125	125	60	54613
GSE25066	508	203	203	102	22283
GSE30784	229	92	92	45	54675

with data fusion, representing an approximately 12% increase in F1-score. On other hand, the average SD value of FedSPL was less than that of most compared methods, which indicates that the proposed method is robust in general. In terms of the F1-score obtained on various data, the result of FedSPL common was better than the compared methods, which implies the effectiveness of FedSPL model is better than other learning methods.

To present the distribution of the measured and modeled performance, boxplots of the measured and estimated performance values are shown in Figure 2. Figure 2 also demonstrates that the performances of FedSPL is greater than the compared methods. In the FedAvg method, data from multiple clients are combined to train a global model using federated learning directly. As shown in Table 2 and Figure 2, the performance of FedAvg was less than personal client in a few scenes. For example, the accuracy and F1-score of FedAvg is less than those of client A on the GSE4115 dataset. This information indicates a few of existed data noise decrease performances of federated learning model.

Discussion

The core components of the proposed FedSPL include the adaptive reweight generation process and the privacy-preserving mechanism. We selected multiple gene expression datasets with different methods to evaluate FedSPL compared with the FedAvg and conventional data fusion methods. Table 2 presents the average classification results obtained by these algorithms in terms of F1-score, accuracy and AUC. The results demonstrate that FedSPL outperformed other competitors on average on all these datasets. This performance improvement is explained as follows.

Table 2. Comparison of performances with different approaches on different datasets

Dataset	Terms	Client A	Client B	FedAvg	Data fusion	FedSPL
GSE4115	Accuracy	0.674±0.08	0.584±0.05	0.655 ±0.06	0.692 ±0.06	0.703±0.03
	AUC	0.731±0.08	0.676±0.05	0.732±0.05	0.764±0.05	0.770±0.04
	Recall	0.675±0.14	0.575±0.11	0.660 ±0.12	0.695 ±0.07	0.698±0.09
	Precision	0.693±0.07	0.610±0.05	0.675±0.05	0.712 ±0.06	0.736±0.05
	F1-score	0.679±0.09	0.586±0.07	0.662 ±0.08	0.702±0.06	0.708 ±0.05
GSE21050	Accuracy	0.831 ±0.04	0.815±0.04	0.826±0.04	0.831 ± 0.05	0.839±0.04
	AUC	0.871 ±0.05	0.868 ±0.04	0.885 ±0.04	0.878 ±0.04	0.886 ±0.05
	Recall	0.911 ±0.05	0.903 ±0.06	0.903 ±0.04	0.914 ±0.03	0.911 ±0.04
	Precision	0.828 ±0.05	0.813 ±0.05	0.825 ±0.04	0.825 ±0.04	0.837 ±0.04
	F1-score	0.865 ±0.03	0.853 ±0.03	0.861 ±0.03	0.866 ±0.03	0.871 ±0.02
GSE25066	Accuracy	0.776 ± 0.01	0.776 ± 0.02	0.764 ± 0.02	0.770 ± 0.01	0.786 ±0.01
	AUC	0.628 ± 0.06	0.605 ±0.07	0.649 ±0.05	0.640 ±0.03	0.654 ±0.05
	Recall	0.076 ±0.03	0.071 ±0.04	0.096 ±0.06	0.045 ±0.01	0.136 ±0.04
	Precision	0.386 ±0.08	0.464 ±0.22	0.330 ±0.14	0.321 ±0.10	0.393 ±0.06
	F1-score	0.126 ± 0.05	0.118 ± 0.05	0.146 ±0.08	0.079 ±0.01	0.199 ±0.05
GSE30784	Accuracy	0.900 ±0.03	0.917± 0.03	0.900± 0.03	0.915±0.02	0.920 ±0.03
	AUC	0.943 ± 0.03	0.936 ± 0.04	0.949±0.02	0.942 ± 0.03	0.950 ± 0.04
	Recall	0.950± 0.02	0.962 ± 0.03	0.947 ±0.04	0.953 ±0.02	0.947 ± 0.02
	Precision	0.919 ±0.03	0.930± 0.03	0.921± 0.02	0.935 ± 0.02	0.945 ± 0.02
	F1-score	0.933 ± 0.02	0.945± 0.02	0.933 ± 0.02	0.943 ± 0.02	0.946 ± 0.02

Table 3. Key genes were selected from GSE4115

Gene ID	Gene symbol	P-value
202732_at	Protein kinase (cAMP-dependent, catalytic) inhibitor gamma (PKIG)	<0.001
205686_s_at	CD86 molecule(CD86)	<0.001
206112_at	Ankyrin repeat domain 7 (ANKRD7)	<0.001
206627_s_at	SSX family member 1 (SSX1)	<0.001
207782_s_at	Presenilin 1 (PSEN1)	<0.001
210167_s_at	TEF, PAR bZIP transcription factor (TEF)	<0.001
211104_s_at	Myosin VIIA (MYO7A)	<0.001
211619_s_at	Alkaline phosphatase, placental (ALPP)	<0.001
212998_x_at	Major histocompatibility complex, class II, DQ beta 1 (HLA-DQB1)	<0.001
214260_at	COP9 signalosome subunit 8 (COPS8)	<0.001
217866_at	Cleavage and polyadenylation specific factor 7 (CPSF7)	<0.001
218163_at	MCTS1, re-initiation and release factor (MCTS1)	<0.001
219232_s_at	egl-9 family hypoxia inducible factor 3 (EGLN3)	<0.001
219457_s_at	Ras and Rab interactor 3 (RIN3)	<0.001
219519_s_at	Sialic acid binding Ig like lectin 1 (SIGLEC1)	<0.001
219956_at	Polypeptide *N*-acetylgalactosaminyltransferase 6 (GALNT6)	<0.001
34187_at	RNA-binding motif single-stranded interacting protein 2 (RBMS2)	<0.001

Table 4. Key genes were selected from GSE21050

Gene ID	Gene symbol	P-value
205875_s_at	Three prime repair exonuclease 1 (TREX1)	<0.001
219872_at	Family with sequence similarity 198 member B (FAM198B)	<0.001
222255_at	Periaxin (PRX)	<0.001
225595_at	CREB/ATF bZIP transcription factor (CREBZF)	<0.001
238520_at	Transcriptional regulating factor 1 (TRERF1)	<0.001
242903_at	Interferon gamma receptor 1 (IFNGR1)	<0.001
242986_at	Neuron navigator 1 (NAV1)	<0.001

1) The adaptive reweight generation process is a pivotal component of FedSPL. The adaptive density-based reweight sample learning mechanism considers noisy and intractable samples to improve model robustness, which utilizes the prior distribution information of the original data to further distinguish differences between samples and minimize the influence of outliers or noises in these data.

Table 5. Key genes was selected from GSE25066

Gene ID	Gene symbol	P-value
201235_s_at	BTG anti-proliferation factor 2 (BTG2)	<0.001
201596_x_at	Keratin 18(KRT18)	<0.001
202633_at	Topoisomerase (DNA) II binding protein 1 (TOPBP1)	<0.001
203566_s_at	Amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase (AGL)	<0.001
203859_s_at	Paralemmin (PALM)	<0.001
203888_at	Thrombomodulin (THBD)	<0.001
204289_at	Aldehyde dehydrogenase 6 family member A1 (ALDH6A1)	<0.001
206380_s_at	Complement factor properdin (CFP)	<0.001
206951_at	Histone cluster 1 H4 family member e(HIST1H4E)	<0.001
214738_s_at	NIMA related kinase 9 (NEK9)	<0.001
216079_at	Epilepsy, progressive myoclonus type 2A, Lafora disease (laforin) (EPM2A)	<0.001
216321_s_at	Nuclear receptor subfamily 3 group C member 1 (NR3C1)	<0.001
218427_at	Serologically defined colon cancer antigen 3 (SDCCAG3)	<0.001
218615_s_at	Transmembrane protein 39A (TMEM39A)	<0.001
220015_at	Castor zinc finger 1 (CASZ1)	<0.001

Table 6. Key genes was selected from GSE30784

Gene ID	Gene symbol	P-value
1007_s_at	microRNA 4640 (MIR4640)	<0.001
1555500_s_at	SLC2A4 regulator (SLC2A4RG)	<0.001
202581_at	Heat shock protein family A (Hsp70) member 1A (HSPA1A)	<0.001
205827_at	Cholecystokinin (CCK)	<0.001
206604_at	Ovo like transcriptional repressor 1 (OVOL1)	<0.001
206655_s_at	SEPT5-GP1BB readthrough (SEPT5-GP1BB)	<0.001
206857_s_at	FK506 binding protein 1B (FKBP1B)	<0.001
209105_at	Nuclear receptor coactivator 1 (NCOA1)	<0.001
210844_x_at	Catenin alpha 1 (CTNNA1)	<0.001
211923_s_at	Zinc finger protein 471 (ZNF471)	<0.001
214363_s_at	Small nucleolar RNA host gene 4 (SNHG4)	<0.001
216949_s_at	Polycystin-1-like (LOC101930075)	<0.001
219621_at	Claspin (CLSPN)	<0.001
220087_at	Beta-carotene oxygenase 1 (BCO1)	<0.001
221751_at	Pantothenate kinase 3 (PANK3)	<0.001
222394_at	Programmed cell death 6 interacting protein (PDCD6IP)	<0.001
222580_at	Zinc finger protein 644 (ZNF644)	<0.001
222691_at	Solute carrier family 35 member B3 (SLC35B3)	<0.001

2) FedSPL obtains the best performance on all four datasets, which indicates the benefit of federated learning because the results of each client after SPL reweights were positive in each iterative process, and our federated learning mechanism enhanced the model's discrimination/classification capability.

However, the local computing complexity of the proposed FedSPL is higher than FedAvg because the self-paced process iterated multiple times to select the high-confidence samples to train the local models. These additional local computing costs are worthwhile to improve the global learning result.

To better understand disease diagnosis based on gene expression data, the significant and highly related disease-causing genes were selected by the embedded regularization approach, as shown in Tables 3, 4, 5 and 6. The *P*-value computed by Pearson's trend test presents the signification of each gene expression. These important genes obtained by the embedded regularization have significant differences according to signification of the *P*-value. Various medical literatures were used to check these gene expressions. For example, a key gene named the CD86 molecule (CD86) was selected from the GSE4115 (lung cancer) gene expression data. After checking with medical experts and querying through the open resource of the National Center for Biotechnology Information[1], we found that CD86 is associated with gene expression in lung disease [55], which encodes a type I membrane protein that is a member of the immunoglobulin superfamily. In addition, the first selected gene in the GSE21050 gene expression data was TREX1. TREX1 plays a key role function in the prevention of host DNA accumulating after cell death, which could actuate an autoimmune response [56]. Similarly, the first selected gene in the GSE25066 (breast cancer) gene expression data was the BTG2, which is a potential biomarker that has been successfully confirmed in breast cancer from previous

[1] https://www.ncbi.nlm.nih.gov/gene/942

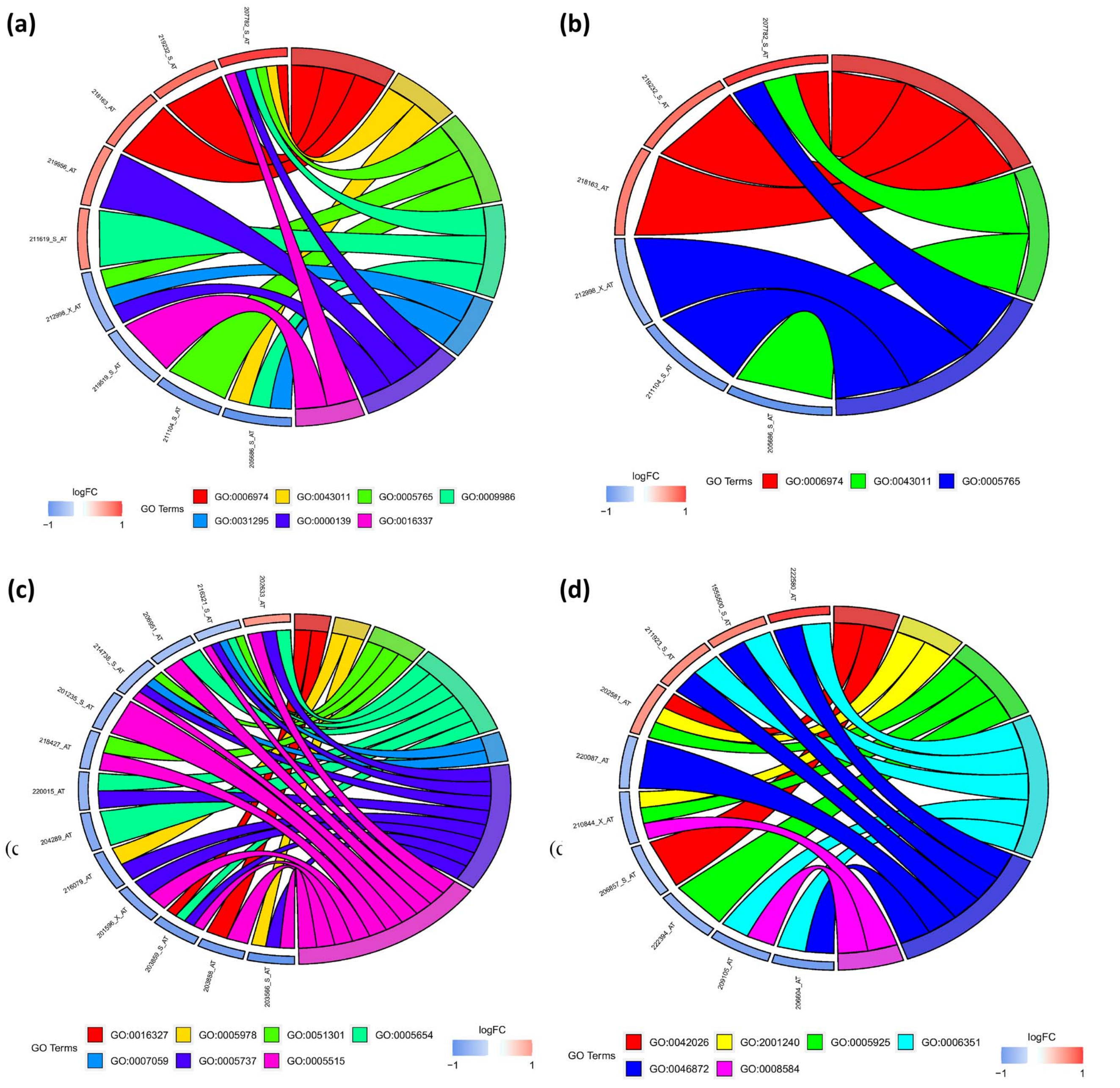


Figure 3. GOChord plot of top overrepresented GO terms. The selected genes are linked to their assigned terms via ribbons, where blue-to-red color squares next to the selected genes indicates the amplitude change - logFC. (a) GSE4115 dataset, (b) GSE21050 dataset, (c) GSE25066 dataset, (d) GSE30784 dataset.

a literature [57]. For the GSE30784 (colorectal cancer) dataset, SLC2A4RG was selected as an important gene here. SLC2A4RG is an important transcription factor involved in SLC2A4 and HD gene transactivation [58]. In summary, existing literature proves the effectiveness of the proposed method, and some other selected high-relative genes can be further investigated through biological experiments in future.

The high-relevant disease causing-genes enriched analysis is shown in Figure 3. It is known key information from Figure 3 that some selected genes have the same pathway, and some genes have a different pathway. For example, GO:0006974 includes three genes in Figure 3**a**. In addition, the colors (blue–red) represent the *LogFC* value, which is shown as up- and downregulation genes, e.g. PSEN1 (gene ID: 207782_S_AT) is an upregulation gene and CD86 (gene ID: 205686_S_AT) is a downregulation gene. In addition, as shown in Figure 3**d**, OVOL1 (gene ID: 206604_AT) is a key gene in oral squamous cell carcinoma, which is a downregulated gene and is a small cluster of EMT-related transcription factors, whose changes are inversely correlated with the tumor EMT phenotype [59].

Conclusion

Date enhancement can improve the performance of the model. Hoverer, lacking labeled samples, overfitting will occur in the learning algorithms used in medical data mining applications, e.g. gene expression data classification, because it is expensive to generate labeled instances.

Thus, we have proposed the FedSPL method to diagnose disease procedures and gene selection in multiple real cancer applications. The proposed method reduces the influence of noisy samples to improve the performance of training model, i.e. removes an apart of irrelevant and redundant instances to obtain a suitable subset. In this modal, model performance is improved and the risk of privacy leaks is reduced. Compared with conventional methods, we found that the FedSPL method not only can protect privacy data but also achieve an improved F1-score of approximately 4.3% (=(((0.708 + 0.871 + 0.199 + 0.946)/4) /((0.662 + 0.861 + 0.156 + 0.933)/4) − 1) × 100%) compared with the mean value of the FedAvg method. To the best of our knowledge, this is the first treatment of self-paced sampling to apply in federated learning problems.

Federated learning has the potential to communicate all the isolated medical institutions and hospitals to share their individual medical data with privacy protocols guarantee. Such collaborative data sharing is expected to greatly improve disease prognosis accuracy in many healthcare applications. In future work, we plan to develop a sophisticated large-scale FedSPL platform by exploring parallel privacy-preserving techniques to closely connect hospitals worldwide and share medical data safely in order to accelerate diseases diagnosis using various large-scale datasets.

Key Points

- A federated self-paced learning named FedSPL is developed with privacy-preserving.
- FedSPL effectively selects samples from high- to low-confidence in each individual client.
- FedSPL can avoids accidental poisoning of the global model from data noises.
- The performance is higher than that of the individual competitors.
- The model selects small and highly relevant genes responsible for early prognosis.

Acknowledgment

The authors would like to thank reviewers for their constructive comments.

Funding

National Natural Science Foundation of China (grant number: 61703416); Training Program for Excellent Young Innovators of Changsha (grant number: KQ2009009); Huxiang Youth Talent Support Program (grant number: 2021RC3076); Postgraduate Scientific Research Innovation Project of Hunan Province (grant number: CX20190040).

References

1. Lian C, Liu M, Zhang J, *et al.* Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**(4):880–93.
2. Wang Y, Bhattacharya T, Jiang Y, *et al.* A novel deep learning method for predictive modeling of microbiome data. *Brief Bioinform* 2021;**22**(3):1–14.
3. Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform* 2021;**22**(1):77–87.
4. Q. Wang, L.-Y. Xia, H. Chai, Y. Zhou, Semi-supervised learning with ensemble self-training for cancer classification, in: 2018 IEEE smart world, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), *IEEE*, **2018**, pp. 796–803.
5. Wang Q, Zhou Y, Ding W, *et al.* Random forest with self-paced bootstrap learning in lung cancer prognosis. *ACM Trans Multimedia Comput Commun Appl (TOMM)* 2020;**16**(1s):1–12.
6. Qi Q, Li Y, Wang J, *et al.* Label-efficient breast cancer histopathological image classification. *IEEE J Biomed Health Inform* 2018;**23**(5):2108–16.
7. Zhong Z, Zheng L, Kang G, *et al.* Random erasing data augmentation. *AAAI* 2020;13001–8.
8. Muzammal M, Talat R, Sodhro AH, *et al.* A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks. *Information Fusion* 2020;**53**:155–64.
9. Zhao X, Jiao Q, Li H, *et al.* ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinformatics* 2020;**21**(1):43.
10. Li P, Li J, Huang Z, *et al.* Multi-key privacy-preserving deep learning in cloud computing. *Fut Gener Comput Syst* 2017;**74**: 76–85.
11. Xing K, Hu C, Yu J, *et al.* Mutual privacy preserving -means clustering in social participatory sensing. *IEEE Trans Industrial Inf* 2017;**13**(4):2066–76.
12. Mohassel P, Rosulek M, Trieu N. Practical privacy-preserving k-means clustering. *IACR Cryptol ePrint Arch* 2020;**2019**:1158.
13. Chang V, Ramachandran M. Towards achieving data security with the cloud computing adoption framework. *IEEE Trans Serv Comput* 2016;**9**(1):138–51.
14. Zhang T, Zhu Q. Dynamic differential privacy for admm-based distributed classification learning. *IEEE Trans Inf Forensics Security* 2017;**12**(1):172–87.
15. Xu J, Glicksberg BS, Su C, *et al.* Federated learning for healthcare informatics. *J Healthcare Inf Res* 2021;**5**(1):1–19.
16. Chen S, Xue D, Chuai G, *et al.* Fl-qsar: a federated learning-based qsar prototype for collaborative drug discovery. *Bioinformatics* 2020;**36**(22–23):5492–8.
17. Warnat-Herresthal S, Schultze H, Shastry KL, *et al.* Swarm learning for decentralized and confidential clinical machine learning. *Nature* 2021;**594**(7862):265–70.
18. Yang D, Xu Z, Li W, *et al.* Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Med Image Anal* 2021;**70**: 101992.
19. Kasyap H, Tripathy S. Privacy-preserving decentralized learning framework for healthcare system, ACM transactions on

multimedia computing. *Commun Appl (TOMM)* 2021;**17**(2s): 1–24.

20. Can YS, Ersoy C. Privacy-preserving federated deep learning for wearable iot-based biomedical monitoring. *ACM Trans Internet Technol (TOIT)* 2021;**21**(1):1–17.
21. Kumar MP, Packer B, Koller D. Self-paced learning for latent variable models, in. *Adv Neural Inf Process Syst* 2010; 1189–97.
22. Wang Q, Zhou Y, Zhang W, *et al.* Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis. *Expert Syst Appl* 2020;**152**:113334.
23. Xia L, Yang Z, Zhang H, *et al.* Improved prediction of drug-target interactions using self-paced learning with collaborative matrix factorization. *J Chem Inf Model* 2019;**59**(7):3340–51.
24. Xia LY, Wang QY, Cao Z, *et al.* Descriptor selection improvements for quantitative structure-activity relationships. *Int J Neural Syst* **29**(9).
25. Chen C, Zhang Q, Ma Q, *et al.* Lightgbm-ppi: predicting protein-protein interactions through lightgbm with multi-information fusion. *Chemom Intel Lab Syst* 2019;**191**:54–64.
26. Chai H, Zhou X, Zhang Z, *et al.* Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput Biol Med* 2021;**134**:104481.
27. Yang P, Xu L, Zhou BB, *et al.* A particle swarm based hybrid system for imbalanced medical data sampling. *BMC Genomics* 2009;**10**(3):1–14.
28. Rozenberg B, Gudes E. Association rules mining in vertically partitioned databases, *Data and knowledge engineering*. 2006;**59**(2):378–96.
29. Liu M, Hu H, Xiang H, *et al.* Clustering-based efficient privacy-preserving face recognition scheme without compromising accuracy. *ACM Trans Sensor Netw (TOSN)* 2021;**17**(3):1–27.
30. Kaissis G, Ziller A, Passerat-Palmbach J, *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat Mach Intell* 2021;**3**(6):473–84.
31. Abadi M, Chu A, Goodfellow I, *et al. Deep learning with differential privacy*. New York, NY: Association for Computing Machinery, 2016.
32. Shokri R, Shmatikov V. *Privacy-preserving deep learning*. New York, NY: Association for Computing Machinery, 2015.
33. Truex S, Baracaldo N, Anwar A, *et al.* A hybrid approach to privacy-preserving federated learning. *Proc 12th ACM Workshop Artif Intell Security* 2019;1–11.
34. Brisimi TS, Chen R, Mela T, *et al.* Federated learning of predictive models from federated electronic health records. *Int J Med Inform* 2018;**112**:59–67.
35. Shen J, Nicolaou CA. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discov Today Technol* 2019;**32**:29–36.
36. Zhou Y, Wang F, Tang J, *et al.* Artificial intelligence in covid-19 drug repurposing. *Lancet Digital Health* 2020;**2**(12):e667–76.
37. Hao M, Li H, Luo X, *et al.* Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Trans Industr Inform* 2020;**16**(10):6532–42.
38. Chen Y, Qin X, Wang J, *et al.* Fedhealth: a federated transfer learning framework for wearable healthcare. *IEEE Intell Syst* 2020;**35**(4):83–93.
39. Tolpegin V, Truex S, Gursoy ME, *et al.* Data poisoning attacks against federated learning systems. *Eur Symp Res Comput Secur Springer* 2020;480–501.
40. McMahan B, Moore E, Ramage D, *et al.* Communication-efficient learning of deep networks from decentralized data. *Artif Intell Stat PMLR* 2017;1273–82.
41. Feki I, Ammar S, Kessentini Y, *et al.* Federated learning for covid-19 screening from chest x-ray images. *Appl Soft Comput* 2021;**106**:107330.
42. Ryffel T, Trask A, Dahl M, *et al.* A generic framework for privacy preserving deep learning. *arXiv preprint* arXiv:1811.04017.
43. Yang Q, Liu Y, Cheng Y, *et al.* Federated learning, synthesis lectures on artificial intelligence and machine. *Learning* 2019;**13**(3):1–207.
44. Wang W, Feng R, Chen J, *et al.* Nodule-plus r-cnn and deep self-paced active learning for 3d instance segmentation of pulmonary nodules. *IEEE Access* 2019;**7**:128796–805.
45. Wang W, Feng R, Liu X, *et al.* Deep active self-paced learning for biomedical image analysis. *Deep Learn Healthcare Springer* 2020;95–110.
46. Yang J, Wu X, Liang J, *et al.* Self-paced balance learning for clinical skin disease recognition. *IEEE Trans Neural Netw Learn Syst* 2019;**31**(8):2832–46.
47. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**(19):2507–17.
48. Golub TR, Slonim DK, Tamayo P, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**(5439):531–7.
49. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;**3**(Mar):1157–82.
50. El Akadi A, Amine A, El Ouardighi A, *et al.* A two-stage gene selection scheme utilizing mrmr filter and ga wrapper. *Knowl Inf Syst* 2011;**26**(3):487–500.
51. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst Appl* 2009;**36**(2): 3240–7.
52. Hua J, Tembe WD, Dougherty ER. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognit* 2009;**42**(3):409–24.
53. Zhang D, Meng D, Zhao L, Han J, Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 3538–44.
54. SHU J, MENG D, XU Z. Meta self-paced learning. *Sci Sinica Inform* 2020;**50**(6):781–93.
55. Tsuyuki S, Tsuyuki J, Einsle K, *et al.* Costimulation through b7-2 (cd86) is required for the induction of a lung mucosal T helper cell 2 (TH2) immune response and altered airway responsiveness. *J Exp Med* 1997;**185**(9):1671–80.
56. Tao S-S, Wu G-C, Zhang Q, *et al.* Trex1 as a potential therapeutic target for autoimmune and inflammatory diseases. *Curr Pharm Des* 2019;**25**(30):3239–47.
57. Kawakubo H, Carey JL, Brachtel EF, *et al.* Expression of the nf-b-responsive gene btg2 is aberrantly regulated in breast cancer. *Oncogene* 2004;**23**(50):8310–9.
58. Dai W, Xu Y, Mo S, *et al.* Glut3 induced by ampk/creb1 axis is key for withstanding energy stress and augments the efficacy of current colorectal cancer therapies. *Signal Transduct Target Ther* 2020;**5**(1):1–14.
59. Zheng X, Wu K, Liao S, *et al.* Microrna-transcription factor network analysis reveals mirnas cooperatively suppress rora in oral squamous cell carcinoma. *Oncogene* 2018;**7**(10):1–18.