Heliyon 9 (2023) e18615

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Predicting anticancer drug sensitivity on distributed data sources using federated deep learning

Xiaolu Xu [a], Zitong Qi [b], Xiumei Han [c], Aiguo Xu [d], Zhaohong Geng [e], Xinyu He [a], Yonggong Ren [a,*], Zhaojun Duo [a,*]

[a] School of Computer and Artificial Intelligence, Liaoning Normal University, Dalian 116029, China
[b] Department of Statistics, University of Washington, Seattle, WA 98195, USA
[c] College of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China
[d] Department of Oncology, The Second People's Hospital of Lianyungang, Lianyungang 222023, China
[e] Department of Cardiology, Second Affiliated Hospital of Dalian Medical University, Dalian 116023, China

ARTICLE INFO

Dataset link: https://1drv.ms/u/s!AoH29XiJLYEbdaJnPQj2vDRRBaw?e=3lggLg

Keywords:
Federated learning
Deep learning
Multi-class focal loss
Drug sensitivity prediction
Gene expression

ABSTRACT

Drug sensitivity prediction plays a crucial role in precision cancer therapy. Collaboration among medical institutions can lead to better performance in drug sensitivity prediction. However, patient privacy and data protection regulation remain a severe impediment to centralized prediction studies. For the first time, we proposed a federated drug sensitivity prediction model with high generalization, combining distributed data sources while protecting private data. Cell lines are first classified into three categories using the waterfall method. Focal loss for solving class imbalance is then embedded into the horizontal federated deep learning framework, i.e., HFDL-fl is presented. Applying HFDL-fl to homogeneous and heterogeneous data, we obtained HFDL-Cross and HFDL-Within. Our comprehensive experiments demonstrated that (i) collaboration by HFDL-fl outperforms private model on local data, (ii) focal loss function can effectively improve model performance to classify cell lines in sensitive and resistant categories, and (iii) HFDL-fl is not significantly affected by data heterogeneity. To summarize, HFDL-fl provides a valuable solution to break down the barriers between medical institutions for privacy-preserving drug sensitivity prediction and therefore facilitates the development of cancer precision medicine and other privacy-related biomedical research.

1. Introduction

Cancer remains an incurable disease worldwide. Cancer genomics studies have shown that each cancer patient possesses a unique genetic mutation profile. Patients with the same type of cancer respond differently to anticancer drugs [1]. Therefore, predicting the clinical response of patients to anticancer drugs based on multiple sources of genomic information and grouping patients for treatment are the focus of research in precision medicine for cancer.

Machine learning (ML) approaches provide powerful tools to predict drug sensitivity in cell lines by mining the relationship between genomic features and drug response metrics rather than time-consuming and expensive wet lab experiments [2]. Jie et al. [3] proposed a deep learning-based efficacy prediction system (DLEPS) that identifies drug candidates using a change in the gene

* Corresponding authors.
E-mail addresses: ygren@lnnu.edu.cn (Y. Ren), duozj609@lnnu.edu.cn (Z. Duo).

https://doi.org/10.1016/j.heliyon.2023.e18615
Received 16 June 2023; Received in revised form 12 July 2023; Accepted 24 July 2023

Available online 1 August 2023

expression profile in the diseased state as input. Validation showed that DLEPS could generate insights into pathogenic mechanisms and drug repurposing. Likun et al. [4] proposed DeepTTA, an end-to-end deep learning model that utilizes a transformer for drug representation learning and a multilayer neural network for transcriptomic data prediction of drug responses. DeepTTA achieved higher performance in terms of root mean square error, Pearson correlation coefficient, and Spearman's rank correlation coefficient on multiple test sets. Ahmed et al. [5] used a graph-based deep learning approach which was evaluated on the Genomics of Drug Sensitivity in Cancer (GDSC) [6] and showed improved predictive performance than the shallow models, e.g., support vector machines and random forest.

However, the decentralization of medicine usually makes the local data volume insufficient for reliable ML model training. Collaboration and data sharing among individual data owners promise to be a good strategy for cost savings and improved predictive performance. However, with the advancement of genomic research, there is a growing privacy concern regarding the collection, storage, and analysis of such sensitive human data for hospitals or research institutions [7–11]. Limitations in the availability of private genomic data have negatively impacted the rate of development of cancer precision medicine based on drug sensitivity studies. Therefore, developing a computational framework for drug sensitivity prediction (DSP) is necessary to share locally sensitive genomic data without compromising private information.

For the first time, Honkela et al. [12] incorporated the differential privacy [13,14] mechanism into Bayesian linear regression for DSP. They balanced the prediction accuracy and privacy protection using 4x more samples compared to non-private regression. However, the proposed method suffers from the curse of dimensionality of gene expression features. Recently, Md. Mohaiminul et al. [15] built a differential privacy deep autoencoder (dpAE) using private gene expression features that performs low-dimensional data representation learning. They extracted GDSC's dpAE features to build a differential privacy DSP model and achieved improved predictive performance than the previous related work. The approaches proposed by Honkela et al. and Md. Mohaiminul et al. both achieve privacy protection for local genomic data but do not take full advantage of data from different institutions to enhance performance for DSP while preserving privacy.

Federated learning (FL) was recently proposed by Google [16–18] and described a distributed and privacy-preserving way of training a global ML model collaboratively without others accessing private data. FL can be divided into horizontal federated learning (HFL), vertical federated learning (VFL), and federated transfer learning (FTL) [19]. HFL is applicable when the datasets share the same feature space but have different samples. VFL is applicable when the datasets share the same sample space but differ in the feature space. FTL is applicable when the two datasets differ not only in the samples but also in the feature space, and only a tiny part of the feature space and the sample space overlap. Three categories of federated learning models are widely used for medical data mining [20–25]. The most common scenario for DSP is that different institutions often have the same feature representations, e.g., gene expression. Compared to VFL and FTL, HFL is more suitable for collaborations among hospitals or research institutions holding genomic data. Essentially, HFL passes encrypted model parameters to the server instead of encrypted raw medical data, which provides a workable solution to the privacy and security issues mentioned above.

In this study, we verified the feasibility of applying HFL to collaborative drug sensitivity prediction based on distributed data sources. Combining the class imbalance property inherent to the prediction, a horizontal federated deep learning model with focal loss function was proposed, denoted as HFDL-fl. In addition to the practical design of the loss, we also studied the performance of federated models for homogeneity and heterogeneity and the parameter aggregation algorithm. We simulated the scenario that parties have their private data respectively and developed HFDL-fl among cross-source (HFDL-Cross) and within-source (HFDL-within). Comprehensive experiments indicated that HFDL-Cross significantly outperforms private deep learning model training on local data. To sum up, our study demonstrated the effectiveness of applying HFL for drug sensitivity prediction for the first time and called for more attention and devotion in this area.

2. Materials and methods

In this study, we proposed a federated learning framework integrating distributed data sources to predict the sensitivity of drugs in cell lines while preserving private information. Fig. 1 shows the detailed pipeline of our proposed framework.

2.1. Data gathering and preprocessing

We integrated two datasets from GDSC [6] and CTRP [26] to build a horizontal federated learning model. The datasets were downloaded by using the R package oncoPredict [27] (https://osf.io/c6tfx/). Sensitivity measure (denoted by $y_{rs,c}$) and gene expression (denoted by $\boldsymbol{x}_{rna,c}$) were considered the response variable and features for cell line c. In the two datasets, some cell lines missed values of the response variable. We removed drugs with 60 or more missing cell line response values. We applied gene expression-based weighted averaging to fill in missing values for the remaining cell lines. Using the weighted averaging method on each dataset, cell lines similar in gene expression space have approximate response values. Detailed steps are as follows:

1. Let z_c^* denote the missing value for the cell line c in the response variable. Let $\boldsymbol{x}_{rna,c}$ denote the vector of gene expression features for the cell line c.
2. Assume cell line k has no missing data for response value. The diversity between cell lines c and k is calculated by $d(c,k) = \|\boldsymbol{x}_{rna,c} - \boldsymbol{x}_{rna,k}\|_2^2$. Search K cell lines nearest to cell line c by calculating the diversity.

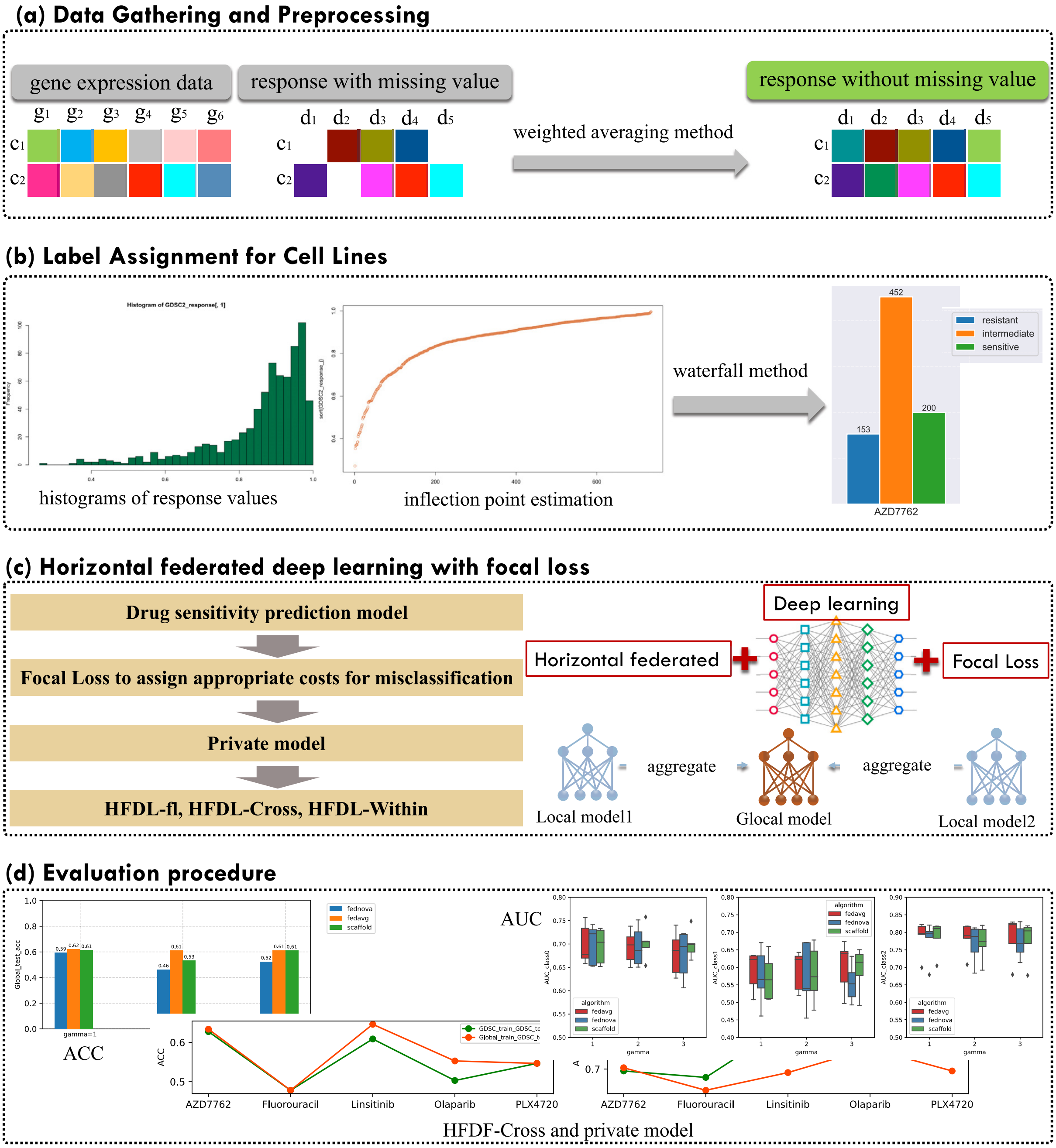
(a) Data Gathering and Preprocessing
gene expression data
g1 g2 g3 g4 g5 g6
c1
c2
response with missing value
d1 d2 d3 d4 d5
weighted averaging method
response without missing value
(b) Label Assignment for Cell Lines
Histogram of GDSC2_response[, 1]
histograms of response values
inflection point estimation
waterfall method
452
153
200
resistant
intermediate
sensitive
AZD7762
(c) Horizontal federated deep learning with focal loss
Drug sensitivity prediction model
Focal Loss to assign appropriate costs for misclassification
Private model
HFDL-fl, HFDL-Cross, HFDL-Within
Deep learning
Horizontal federated
Focal Loss
aggregate
aggregate
Local model1
Glocal model
Local model2
(d) Evaluation procedure
fednova
fedavg
scaffold
Global_test_acc
gamma=1
ACC
AUC
AUC_class0
AUC_class1
AUC_class2
gamma
algorithm
GDSC_train_GDSC_te
Global_train_GDSC_t
AZD7762
Fluorouracil
Linsitinib
Olaparib
PLX4720
HFDF-Cross and private model

Fig. 1. The details pipeline of our proposed framework. (a) Data gathering and preprocessing. (b) Label assignment for cell lines according to response value. (c) Horizontal federated deep learning with focal loss, HFDL-fl, HFDL-Cross, HFDL-Within. (d) Evaluation procedure.

3. Then z_c^* is compensated by

$$\hat{z}_c^* = \sum_{k=1}^{K} \frac{\frac{1}{d(c,k)}}{\sum_{k=1}^{K} \frac{1}{d(c,k)}} z_k.$$

We set $K = 100$ for the preprocessing of GDSC and CTRP datasets. Note that the weighted averaging method is applied separately to each dataset. Specifically, the nearest neighbor cell lines in terms of gene expression values are identified using only those cell lines that belong to the same dataset.

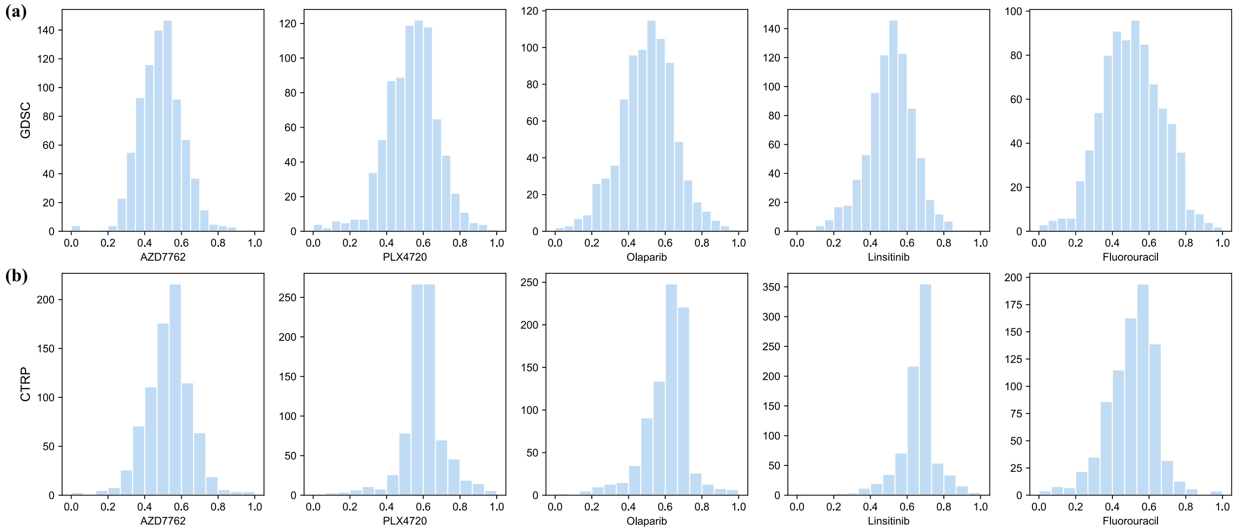

Fig. 2. Histograms of response values of five drugs in both (a) GDSC and (b) CTRP.

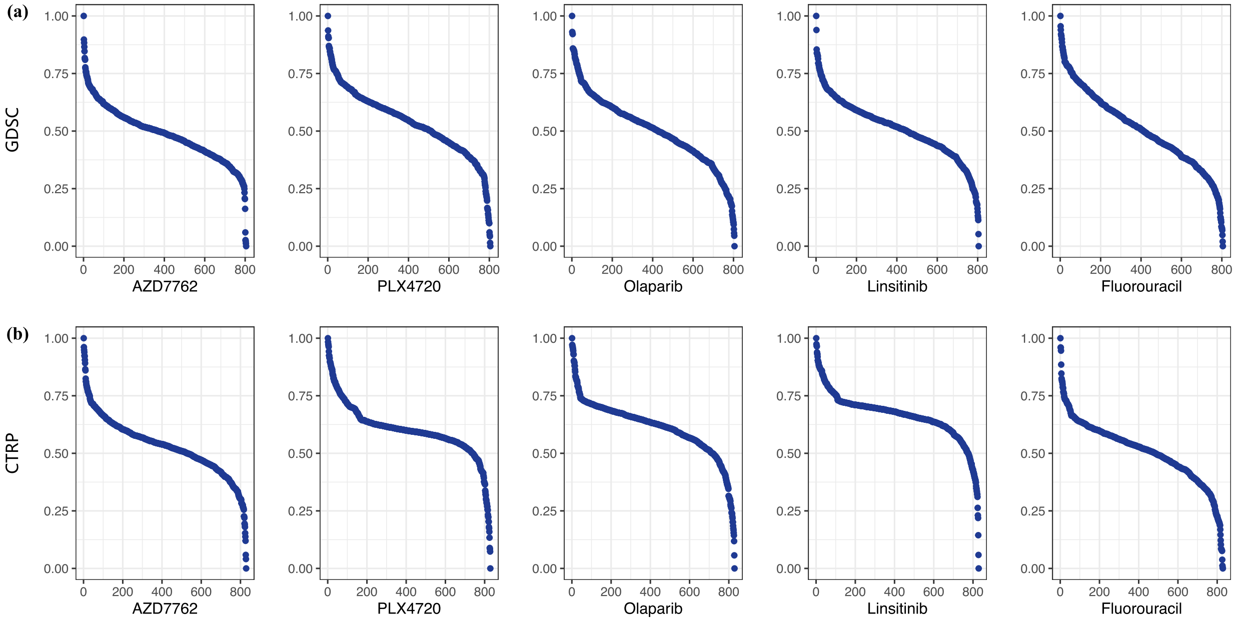

Fig. 3. Waterfall distribution of response values for five drugs in both (a) GDSC and (b) CTRP.

2.2. *Label assignment for cell lines according to response value*

A low response value indicates that the cell line is sensitive to the drug; conversely, a high response value indicates that the cell line is resistant to the drug. Histograms of response values of five drugs in both GDSC and CTRP are shown in Fig. 2. Using the waterfall method, we discretized the drug sensitivity measures into three categories, resistant, intermediate and sensitive [28,29]. The complete procedure is described below:

1. The drug sensitivity measurements, which are the logarithm of IC50 in GDSC and the AUC in CTRP, were extracted and uniformly recorded as RS after being normalized to a range of 0 to 1.
2. Generate a waterfall distribution of RS values, i.e., sort all cell lines based on their RS values in descending order, as shown in Fig. 3.
3. If the waterfall distribution is non-linear (Pearson correlation coefficient to the linear fit ≤ 0.85), estimate the major inflection point of RS curve as the point on the curve with the maximal distance to a line drawn between the start and end points of the distribution.

4. If the waterfall distribution appears linear (Pearson correlation coefficient to the linear fit > 0.85), then use the median RS instead of inflection point.
5. Cell lines within an α-fold (1.2-fold) difference centered around the inflection point are classified as being intermediate, cell lines with lower RS values than this range are defined as sensitive, and those with RS values higher than this range are called resistant.
6. Require at least 5% sensitive and resistant cell lines after applying these criteria.

2.3. Horizontal federated deep learning with focal loss applied in drug sensitivity prediction

We introduced the HFL framework aggregated deep learning model with focal loss as an alternative strategy for collaborative and privacy-preserving drug sensitivity prediction (DSP) modeling. In this section, we describe the pipeline step by step.

2.3.1. Drug sensitivity prediction model

DSP is a triple classification problem with resistant, intermediate, and sensitive categories. Features for cell line k (denoted by $x_k, k = 1, 2, \cdots, N$) are drawn from a feature space $\mathbb{X}$. The corresponding label y_k ($k = 1, 2, \cdots, N$) is drawn from the label space $\mathbb{Y} := \{1, 2, 3\}$ (resistant, intermediate, and sensitive). Let the features corresponding to resistant, intermediate, and sensitive are denoted $\mathbb{X}_r$, $\mathbb{X}_i$, $\mathbb{X}_s$. That are

$$\mathbb{X}_r = \{x_k \in \mathbb{X} : \ y_k = 1\}, \mathbb{X}_i = \{x_k \in \mathbb{X} : \ y_k = 2\}, \text{and } \mathbb{X}_s = \{x_k \in \mathbb{X} : \ y_k = 3\}.$$

For any $x_k^r \in \mathbb{X}_r$, $x_k^i \in \mathbb{X}_i$, and $x_k^s \in \mathbb{X}_s$, the objective is to construct a function $f : \mathbb{X} \to \mathbb{Y}$ such that

$$f(x_k^r) = 1, f(x_k^i) = 2, f(x_k^s) = 3.$$

2.3.2. Focal loss to assign appropriate costs for class imbalance

Class imbalance is intrinsic to DSP problem. Since most classification algorithms assume balanced class distributions and assign equal misclassification costs. They fail to represent the characteristics of imbalanced class and are more likely to classify new samples to the majority class [30]. For DSP, the costs of false resistant and sensitive classifications should be much higher than that of intermediate. To mitigate the challenge of skewed class distribution, we incorporated focal loss [31] to increase the cost associated with misclassifying resistant or sensitive samples.

For the DSP problem, the focus loss function is given by:

$$L_{FL} = -\sum_{t=1}^{C} \alpha_t (1 - p_t)^{\gamma} y_t \log p_t, \tag{1}$$

where C denotes the number of categories ($C = 3$ in this study), p_t denotes a probability distribution of the prediction, α_t denotes the weight factor which down-weights easy samples, y_t denotes a real probability distribution. As shown in equation (2), where $y_t = 1$ if t belongs to the true label, else $y_t = 0$.

$$y_t = \begin{cases} 1 & (t = \text{true label}) \\ 0 & (t \neq \text{true label}) \end{cases}. \tag{2}$$

In equation (1), $\alpha_t(1 - p_t)^{\gamma}$ is added to the cross entropy loss, and the focal loss is obtained. This way, the loss function will focus training on minority samples (resistant or sensitive). Two parameters affect the action of the focal loss on classification, α_t and focusing parameter γ. Parameter α_t for each category is generally set to the inverse of the sample proportion in binary classification, but it does not work in the multiclassification problem in this study. Besides, the focusing parameter γ is used to control the extent to which easy examples are down-weighted. In particular, L_{FL} degenerates to the cross entropy loss when $\gamma = 0$. For $\gamma > 0$, the higher the γ, the greater the effect of modulating factor $(1 - p_t)^{\gamma}$ on the loss. If a sample is correctly classified, p_t is close to 1, $(1 - p_t)^{\gamma}$ tends to 0, and the loss for the sample will decrease significantly. In contrast, if a sample is misclassified, p_t will be small, $(1 - p_t)^{\gamma}$ tends to 1, and the loss remains essentially constant. Overall, it is equivalent to increasing the weight of inaccurately classified samples in the loss. We adopt the multi-class focal loss with $\gamma = 1, 2, 3$ in our experiments.

2.3.3. Private model

Since data owners may not want to publicly release genomic datasets for privacy concerns. In such cases, a site has to only rely on its local data for predictive analytics. We designed a private model of DSP for this scenario, which trains the classifier on local data source. Initially, we did a train test split on the whole dataset: 80% for training and 20% for testing, respectively, for each site. Then, we standardized the features of both the training and testing sets. As a result of this preprocessing, we obtained $\mathbb{X}_{train}$ and $\mathbb{Y}_{train}$, which represent the feature and label sets for training, and $\mathbb{X}_{test}$ and $\mathbb{Y}_{test}$, which represent the feature and label sets for testing. The proportions of the three categories in training and testing sets are the same as the entire dataset using stratified sampling. We used two data sources, GDSC and CTRP, to train the private model. The corresponding experimental data are recorded as $\mathbb{X}_{train\_GDSC}$, $\mathbb{Y}_{train\_GDSC}$, $\mathbb{X}_{test\_GDSC}$, $\mathbb{Y}_{test\_GDSC}$, and $\mathbb{X}_{train\_CTRP}$, $\mathbb{Y}_{train\_CTRP}$, $\mathbb{X}_{test\_CTRP}$, $\mathbb{Y}_{test\_CTRP}$. Classifier of private model is trained on $\mathbb{X}_{train}$ and $\mathbb{Y}_{train}$, and tested on $\mathbb{X}_{test}$.

2.3.4. Horizontal federated deep learning model with focal loss

We proposed a horizontal federal deep learning model with focal loss (HFDL-fl) across two sources, GDSC and CTRP, to predict drug sensitivity. In the federated learning paradigm, each data owner is denoted as a client, which trains a local model with the same structure. The global model is obtained by aggregating local models. Let T denote the number of rounds for aggregating local model updates. For stochastic gradient descent applied in deep neural network parameter learning, let η, E, B_{train}, and B_{test} denote the learning rate, number of epochs, batch size in training, and batch size in testing, respectively. During local model training, based on given η, E, B_{train}, and B_{test}, gradient for its current model parameter w was computed. We obtained the global model by aggregating parameter updates from the local models by FedAvg [32], FedNova [33], and SCAFFOLD [34]. The process was repeated until round t reached the preset number T. The global model only relies on updates from the local models rather than raw data residing at clients. Algorithm 1 presents the core algorithm of HFDL-fl with FedAvg as the aggregating algorithm.

Algorithm 1: HFDL-fl with FedAvg.

```
Input:
Local datasets for training: $\mathbb{X}_{train\_GDSC}$, $\mathbb{Y}_{train\_GDSC}$, $\mathbb{X}_{train\_CTRP}$, $\mathbb{Y}_{train\_CTRP}$ ;
Parameters in deep learning: learning rate $\eta$, number of epochs $E$, batch size in training $B_{train}$;
Number of communication rounds $T$
Output:
HFDL-fl model parameters for drug sensitivity prediction $w^T$
Server executes:
initialize $x^0$
for $t = 0, 1, \cdots, T-1$ do
    for $i \in (1, 2)$ do
        send the global model $w^t$ to client $P_i$
        $\triangle w_i^t \leftarrow$ LocalTraining($i$, $w^t$)
    end
    $w^{t+1} \leftarrow w^t - \eta \sum_{i \in S_t} \frac{|D^i|}{n} \triangle w_k^t (D^1 = \mathbb{X}_{train\_GDSC}; D^2 = \mathbb{X}_{train\_CTRP})$
end
return $w^T$
Client executes:
$L(w; \boldsymbol{b}) = \sum_{(x,y) \in \boldsymbol{b}} l(w; x; y)$ (focal loss)
LocalTraining($i, w^t$):
$w_i^t \leftarrow w^t$
for epoch $k = 1, 2, \cdots, E$ do
    for each epoch $\boldsymbol{b} = \{\boldsymbol{x}, y\}$ of $D^i$ do
        $w_i^t \leftarrow w_i^t - \eta \triangle L(w_i^t; \boldsymbol{b})$
    end
end
$\triangle w_i^t \leftarrow w^t - w_i^t$
return $\triangle w_i^t$
```

To verify the effect of homogeneity and heterogeneity of client-side data on HFDL-fl, we developed the HFDL-fl model for cross-source (HFDL-Cross) and the HFDL-fl model for within-source (HFDL-within). HFDL-Cross implements horizontal federal deep learning using GDSC as one client and CTRP as another client. The training data of the local models in the two clients is the training data in the private model in Section 2.3.3 ($\mathbb{X}_{train\_GDSC}$ and $\mathbb{X}_{train\_CTRP}$) for a fair comparative analysis. The testing data of the global model in HFDL-Cross is the combination of testing data in two private models ($\mathbb{X}_{test\_GDSC}$ and $\mathbb{X}_{test\_CTRP}$). HFDL-Within implements horizontal federal deep learning within one dataset (GDSC or CTRP). Take HFDL-within in GDSC as an example, we partitioned $\mathbb{X}_{train\_GDSC}$ (training data in the private model) into mutually exclusive sets $\left\{\mathbb{X}^i_{train\_GDSC}\right\}_{i=1}^{N}$ ($N = 2$ in this study), i.e., $\mathbb{X}^1_{train\_GDSC} \cup \mathbb{X}^2_{train\_GDSC} = \mathbb{X}_{train\_GDSC}$ and $\mathbb{X}^1_{train\_GDSC} \cap \mathbb{X}^2_{train\_GDSC} = \emptyset$. We followed the same approach to partition the corresponding label set $\mathbb{Y}_{train\_GDSC}$ into $\mathbb{Y}^1_{train\_GDSC}$ and $\mathbb{Y}^2_{train\_GDSC}$. The global model of HFDL-Within on GDSC will be tested on $\mathbb{X}_{test\_GDSC}$ (testing data in private model). The detailed scheme for the private model, HFDL-Within, and HFDL-Cross, is shown in Fig. 4.

2.4. Evaluation procedure

Accuracy (ACC), AUC_micro, AUC_macro, AUC for class 0 (AUC_class0), AUC for class 1 (AUC_class1) and AUC for class 2 (AUC_class2) [35,36] were selected to evaluate the predictive performance for the private model, HFDL-within, and HFDL-Cross. For classifier $f : D \to C = \{1, \cdots, m\}$ and finite set $S \subset D \times C$, let $a^{f,S} \in \mathbb{N}_0^{m\times m}$ be a confusion matrix, where $a_{ij}^{f,S} = |\{s \in S | f(s_1) = i \wedge s_2 = j\}|$. For the definition of $a_{ij}^{f,S}$, s_1 corresponds to the features of the sample, and s_2 corresponds to the label of the same sample. The Accuracy (ACC) were defined as:

$$\text{accuracy:} \quad ACC = \frac{\sum_x a_{xx}}{\sum_{x=1}^{m} \sum_{i=1}^{m} a_{xi}}.$$

The micro-AUC and macro-AUC are two kinds of the area under the Receiver Operating Characteristic (ROC) curve. Therein, micro calculates metrics globally by considering each element of the label indicator matrix as a label. Macro calculates metrics for each

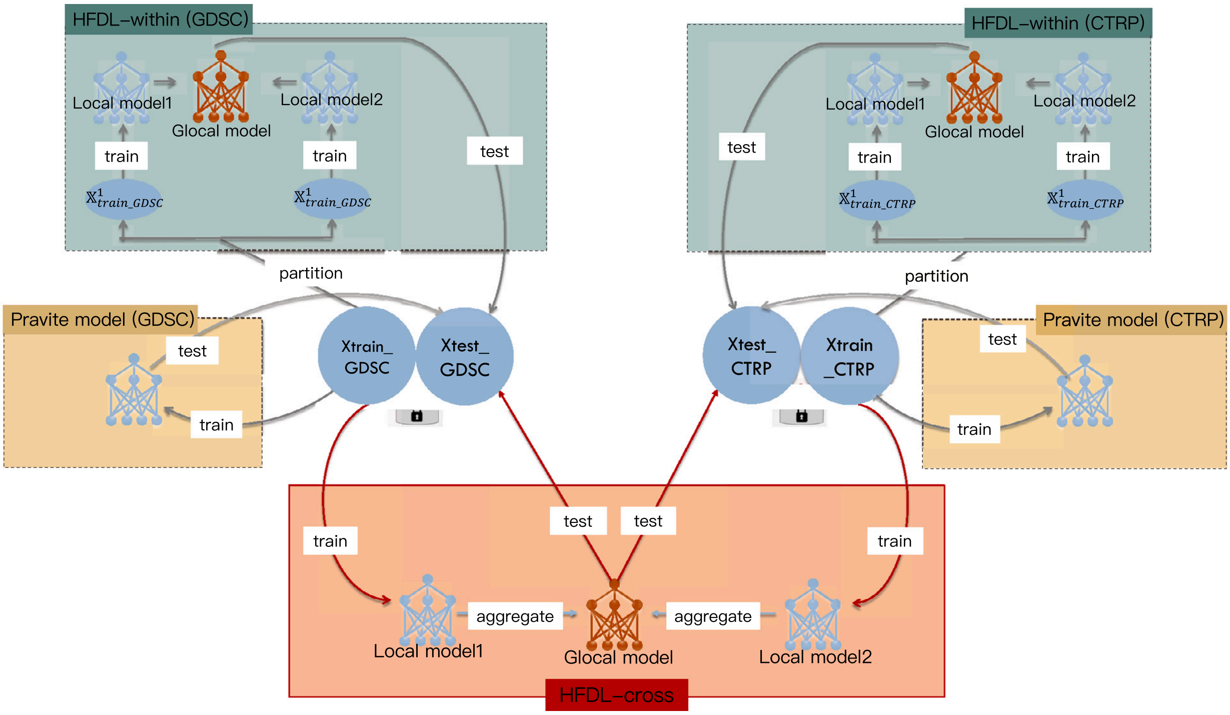

Fig. 4. Scheme of detailed data training and testing for the private model, HFDL-Within, and HFDL-Cross.

Table 1
Total numbers of samples and gene expression features in GDSC and CTRP.

Database	State	Cell lines	Gene expression features
GDSC	Raw	805	17419
	Preprocessed	805	17180
CTRP	Raw	829	51847
	Preprocessed	829	17180

label and finds their unweighted mean, which does not consider label imbalance. The AUC_class0, AUC_class1, and AUC_class2 are the results obtained by calculating the area under the ROC curve for each of the three classes separately. In the calculation process, each class is treated as a positive instance, while the other two classes are treated as negative instances.

3. Results

3.1. Data description

Five drugs, AZD7762, PLX4720, Olaparib, Linsitinib, and Fluorouracil, in both GDSC and CTRP databases, were studied to validate the proposed HFDL-Within and HFDL-Cross. Response data containing missing values were compensated by using the weighted averaging method described in section 2.1. The total number of samples and gene expression features are listed in Table 1. Among them, gene signatures that are common to both databases were preserved. According to the distribution of response values, cell lines in five drugs were classified into three categories: resistant, intermediate, and sensitive. The number of samples in each category is shown in Fig. 5. For Olaparib and Linsitinib in CTRP, the sample size of resistant category is relatively small. The proportions of samples in the three categories are approximately equal in both databases for the other three drugs.

3.2. Performance of HFDL-cross

The overall predictive performance of HFDL-Cross is significantly improved upon private models on local data. Overall, the predictive performance is dependent on two factors: the parameter aggregation algorithms in the global federated model and gamma of focal loss function in deep learning. The detailed performance evaluation process is as follows.

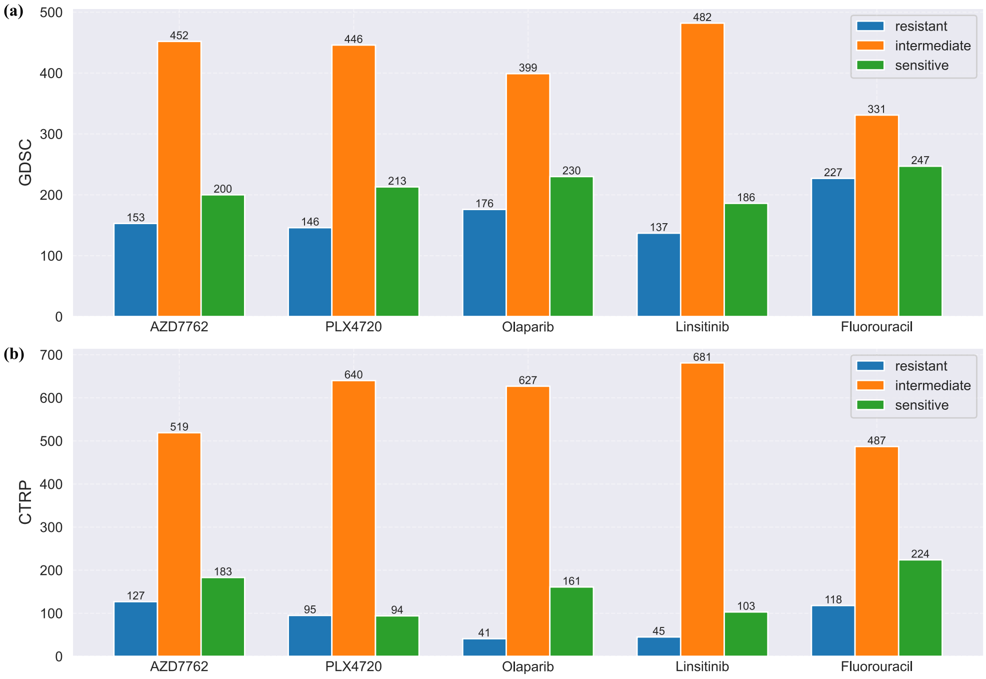

Fig. 5. Number of samples in the three categories in (a) GDSC and (b) CTRP.

Table 2
Performance of HFDL-Cross on five drugs by adjusting gamma and parameter aggregation algorithm.

Drug	Gamma	Algorithm	Global test ACC	AUC_micro	AUC_macro	AUC_class0	AUC_class1	AUC_class2
AZD7762	2	SCAFFOLD	0.6697	0.8037	0.7440	0.7578	0.6464	0.8200
PLX4720	3	FedAvg	0.6208	0.7602	0.6249	0.6271	0.5582	0.6793
Olaparib	1	FedAvg	0.6667	0.7976	0.7111	0.6699	0.6327	0.8225
Linsitinib	2	FedAvg	0.6667	0.8137	0.7047	0.6982	0.6227	0.7829
Fluorouracil	3	FedNova	0.5505	0.7282	0.6552	0.7221	0.4928	0.7436

3.2.1. Overall drug sensitivity predictive performance of HFDL-cross

We applied HFDL-Cross on GDSC (denoted as client 0) and CTRP (denoted as client 1) database. We summarized the predictive performance on five drugs for gamma = 1,2,3 and three aggregation algorithms FedAvg, FedNova, and SCAFFOLD. The results of the round with the highest test ACC were output. The distributions of the six performance criteria on five drugs are shown in Fig. 6. Generally, the overall performance with gamma = 1,2,3 does not show significant performance differences. The aggregation algorithm FedAvg got better performance than SCAFFOLD and FedNova. Besides, the variance of FedNova in the performance of the five drug sensitivity predictions is higher than FedAvg and SCAFFOLD. The average values of the six performance criteria among the five drugs are shown in Fig. 7. The best performance on five drugs by adjusting gamma and aggregation algorithm is shown in Table 2. The best average global test ACC (gamma = 1 and FedAvg algorithm), AUC_micro (gamma = 1 and FedAvg algorithm), AUC_macro (gamma = 1 and FedAvg algorithm), AUC_class0 (gamma = 2 and SCAFFOLD algorithm), AUC_class1 (gamma = 3 and FedAvg algorithm), AUC_class2 (gamma = 1 and SCAFFOLD algorithm) are 0.6196, 0.7770, 0.6946, 0.7030, 0.6027, and 0.7865 respectively. According to the results, the proposed algorithm under the effect of focal loss has relatively better performance on the cell lines that are resistant and sensitive.

3.2.2. Comparison between HFDL-cross and private model

We compared the predictive performance between the private model and HFDL-Cross on the five drugs mentioned above (Fig. 8). In a "small distribution gap" scenario (AZD7762 in Fig. 5), cross-source data collaboration by HFDL-fl achieved better performance than that in "large distribution gap" scenarios (Fluorouracil in Fig. 5). The average ACC, AUC_micro, AUC_macro, AUC_class0, AUC_class1, and AUC_class2 of the private model among five drugs on GDSC are 0.5528, 0.6171, 0.5269, 0.5027, 0.5554 and on CTRP are 0.7554, 0.6548, 0.5542, 0.5627, 0.5275, and 0.5508 respectively. For HFDL-Cross, the results are 0.5714, 0.7360, 0.6803, 0.7205, 0.5512, 0.7542 on GDSC and 0.6963, 0.8224, 0.7094, 0.6613, 0.6442, and 0.8018 on CTRP, respectively. In summary, collaboration by HFDL-fl gained a substantial performance improvement to that of the private model on the local data. More detailed results of HFDL-Cross and private model can be found in Supplementary Tables S1 and S2.

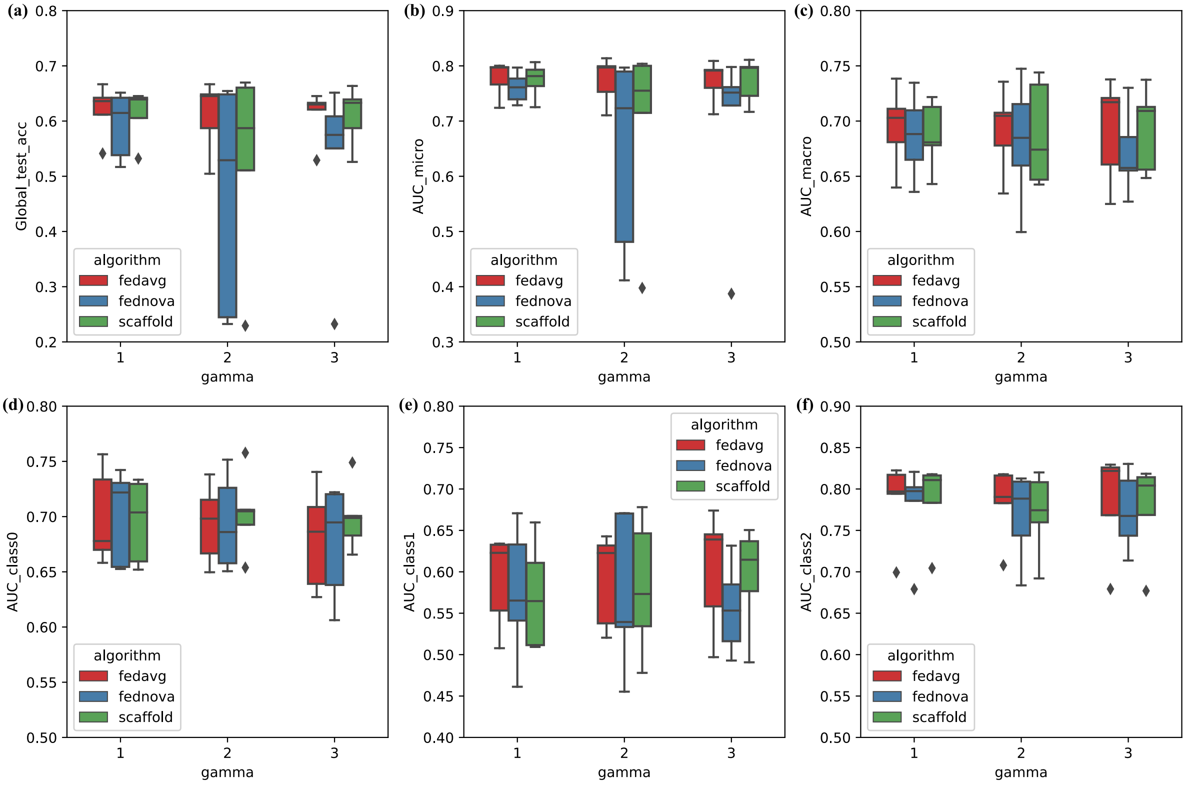

Fig. 6. The distributions of (a) global test ACC, (b) AUC_micro, (c) AUC_macro, (d) AUC_class0, (e) AUC_class1, and (f) AUC_class2 on five drugs.

Table 3
Performance of HFDL-Within by adjusting gamma and parameter aggregation algorithm.

Dataset	Drug	Gamma	Algorithm	Global test ACC	AUC_micro	AUC_macro	AUC_class0	AUC_class1	AUC_class2
GDSC	AZD7762	2	FedAvg	0.6770	0.7478	0.6823	0.6165	0.6558	0.7589
	PLX4720	1	FedNova	0.5466	0.7246	0.6381	0.6802	0.5705	0.6483
	Olaparib	1	FedNova	0.5652	0.7022	0.6913	0.6700	0.5984	0.7937
	Linsitinib	3	FedAvg	0.5466	0.7246	0.6589	0.7631	0.4738	0.7249
	Fluorouracil	1	FedAvg	0.4720	0.5769	0.5676	0.6846	0.4670	0.5369
CTRP	AZD7762	1	FedAvg	0.6988	0.7763	0.7077	0.6767	0.5832	0.8490
	PLX4720	1	SCAFFOLD	0.7470	0.8454	0.6600	0.5882	0.6718	0.7008
	Olaparib	3	FedAvg	0.8133	0.8895	0.7689	0.6313	0.8048	0.8540
	Linsitinib	2	FedAvg	0.7530	0.8773	0.6744	0.5741	0.6150	0.8046
	Fluorouracil	2	SCAFFOLD	0.6566	0.7761	0.6900	0.7200	0.5473	0.7878

3.3. *Performance of HFDL-within*

To access the applicability of HFDL-fl in the homogeneity scenario, we applied HFDL-fl within GDSC or CTRP, respectively, i.e., HFDL-Within. HFDL-Within simulates the situation where samples of clients are homologous or have a similar distribution. Samples in GDSC or CTRP datasets were first divided into training and testing sets. The training set was then divided into two sets as two simulated clients. Similar to the statistics in Section 3.2.1, the results of HFDL-Within modeled on GDSC and CTRP are shown in Table 3. Although data distribution in each client is homogeneous, the performance of HFDL-Within is not significantly improved than that of HFDL-Cross due to the reduced sample size. In addition, HFDL-Within on CTRP shows better performance than HFDL-Within on GDSC among the five drugs. The experimental results revealed that the performance of HFDL-Within did not vary significantly when we used various gamma values of 1, 2, and 3. This finding is in line with the conclusion drawn from the results of HFDL-Cross. Furthermore, FedAvg demonstrated a more stable aggregation performance for the global model.

4. Discussion

Machine learning models can exhibit excellent performance if large-scale public data is available. However, non-restricted sufficient data is always unavailable due to privacy regulations in realistic clinical scenarios. Drug sensitivity prediction studies based

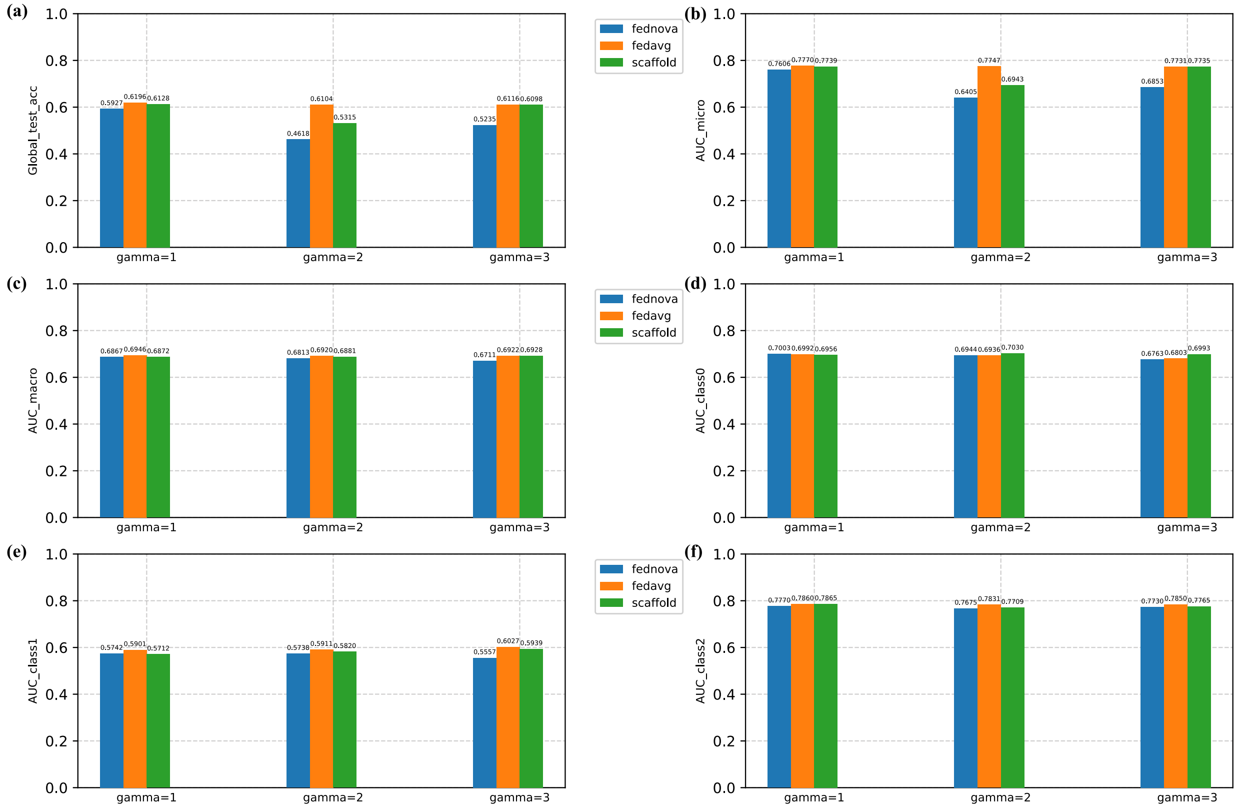

Fig. 7. Average performance for (a) global test ACC, (b) AUC_micro, (c) AUC_macro, (d) AUC_class0, (e) AUC_class1, and (f) AUC_class2 of HFDL-Cross on five drugs.

on machine learning models are deeply plagued by the same problem of limited samples in local hospitals or research institutions, which reduces their clinical application to a certain extent. Federated learning overcomes such challenges by shifting the centralized training approach to the cryptographic delivery of model parameters. Therefore, we proposed a horizontal federated deep learning model with focal loss (HFDL-fl) to improve the generalization of drug sensitivity prediction with privacy preservation. Applying the HFDL-fl model to single and distributed data sources, respectively, we obtained HFDL-Within and HFDL-Cross. We explored three challenging situations for federated learning-based drug sensitivity prediction, loss function design for class imbalance problem (i), the performance of federated learning for homogeneity and heterogeneity scenarios (ii), and parameter aggregation algorithms for global model (iii). To this end, we designed experiments to assess ACC, AUC_micro, AUC_macro, AUC_class0, AUC_class1, and AUC_class2 of HFDL-Cross, HFDL-Within, and private models.

The design of the loss function for class imbalance problem (i) is crucial for both traditional deep learning and federated deep learning. We classified the cell lines into three categories based on IC50 and AUC distribution using the waterfall method (Section 2.2), which effectively avoids the unsmooth excess of dichotomous classification for intermediate category. The ratio of samples in three categories of resistant, intermediate, and sensitive is approximately 1:5:1 (Fig. 5). If misclassifications in every category are assigned the same penalty, federated deep learning model will tend to classify samples as intermediate. The results suggest that despite the proposed algorithm having some misclassification, it attained a high AUC score, particularly for sensitive and resistant categories. This indicates that the classifier has strong discriminative power for accurately distinguishing between these categories. Besides, it can be seen that focal loss can play a more influential role for federated model based on the superior performance of federated model over private model (Fig. 8).

For evaluating the impact of homogeneous and heterogeneous data on federated learning performance (ii), we proposed HFDL-Within and HFDL-Cross. The HFDL-Cross unites two heterogeneous data sources (GDSC and CTRP) and jointly trains about 1000 samples that respond to the same drug. HFDL-Within is modeled on GDSC or CTRP data sources alone, i.e., the homogeneous data from the same data source is divided into two clients for federated training. Overall, HFDL-Cross has better predictive performance than HFDL-Within (Table 2, Table 3, and Fig. 6). HFDL-Cross is only inferior to HFDL-Within on CTRP in terms of ACC. Although homogeneous data are more helpful for parameter aggregation, the reduced sample size will affect the federated model performance of HFDL-Within.

Federated learning can aggregate data across clients to train a joint model whose performance will be influenced by the parameter aggregation algorithm (iii). To obtain the global model, we applied three typical algorithms, FedAvg, SCAFFOLD, and FedNova. Results revealed that FedAvg algorithm exhibited a more stable aggregation performance in both the federated learning for HFDL-

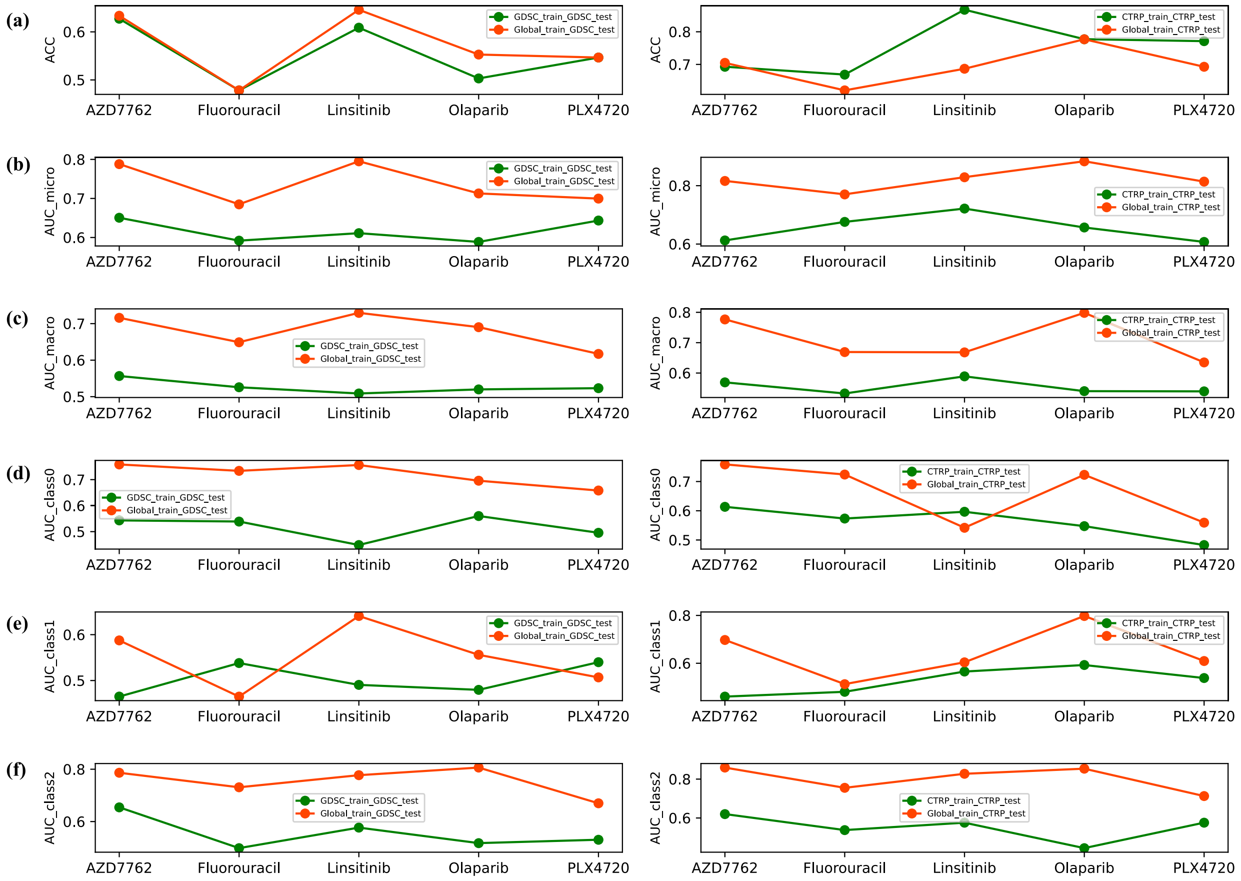

Fig. 8. Comparison between private models and HFDL-Cross on (a) global test ACC, (b) AUC_micro, (c) AUC_macro, (d) AUC_class0, (e) AUC_class1, and (f) AUC_class2.

Cross and HFDL-Within (Table 2 and Table 3). A more detailed comprehensive evaluation and summary of the typical aggregation algorithms can be referred to [37].

Using the conclusions from (i)(ii)(iii), we obtained HFDL-fl across distributed data sources with optimal gamma and aggregation algorithms. Experimental results on five drugs common to both GDSC and CTRP show that our proposed HFDL-Cross outperformed the private model (Fig. 8). The private model exhibited a small performance advantage over HFDL-Cross on AUC_class1 solely for the drug Fluorouracil in GDSC. Furthermore, for the drugs Fluorouracil and Linsitinib in CTRP, the private model has shown a minor improvement over HFDL-Cross on ACC. However, after further examination of the AUC, it was discovered that the private model's high performance for these two predictions compromises AUC. On the other hand, the HFDL-Cross model outperformed the private model in AUC performance for all drugs in both databases.

Overall, the proposed framework of HFDL-fl achieves improved generalizability for drug sensitivity prediction by integrating data from distributed sources while guaranteeing individuals' privacy. We believe that the medical aid platform based on the federated ML-model can provide more accurate analysis and prediction of medical conditions under data security and provide a scientific guarantee for the subsequent promotion of precision medicine.

Funding

The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (No. 61976109, No. 62006108), Liaoning Revitalization Talents Program (No. XLYC2006005), Scientific Research Project of Liaoning Province (No. LJKZ0963), Key R&D projects of Liaoning Provincial Department of Science and Technology, Liaoning Provincial Key Laboratory Special Fund.

CRediT authorship contribution statement

Xiaolu Xu: Conceived and designed the experiments; Performed the experiments, Wrote the paper. **Zitong Qi:** Conceived and designed the experiments; Wrote the paper. **Xiumei Han:** Performed the experiments; Analyzed and interpreted the data. **Aiguo Xu:** Contributed reagents, materials, analysis tools or data. **Zhaohong Geng:** Contributed reagents, materials, analysis tools or data.

Xinyu He: Analyzed and interpreted the data. **Yonggong Ren:** Analyzed and interpreted the data, Wrote the paper. **Zhaojun Duo:** Performed the experiments, Wrote the paper.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data availability

Data associated with this study has been deposited at https://1drv.ms/u/s!AoH29XiJLYEbdaJnPQj2vDRRBaw?e=3lggLg.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e18615.

References

[1] J. Chen, L. Zhang, A survey and systematic assessment of computational methods for drug response prediction, Brief. Bioinform. 22 (1) (2021) 232–246.

[2] L. Parca, G. Pepe, M. Pietrosanto, G. Galvan, L. Galli, A. Palmeri, M. Sciandrone, F. Ferrè, G. Ausiello, M. Helmer-Citterich, Modeling cancer drug response through drug-specific informative genes, Sci. Rep. 9 (1) (2019) 1–11.

[3] J. Zhu, J. Wang, X. Wang, M. Gao, B. Guo, M. Gao, J. Liu, Y. Yu, L. Wang, W. Kong, et al., Prediction of drug efficacy from transcriptional profiles with deep learning, Nat. Biotechnol. 39 (11) (2021) 1444–1452.

[4] L. Jiang, C. Jiang, X. Yu, R. Fu, S. Jin, X. Liu, Deeptta: a transformer-based model for predicting cancer drug response, Brief. Bioinform. 23 (3) (2022) bbac100.

[5] K.T. Ahmed, S. Park, Q. Jiang, Y. Yeu, T. Hwang, W. Zhang, Network-based drug sensitivity prediction, BMC Med. Genom. 13 (11) (2020) 1–10.

[6] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, et al., Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (D1) (2012) D955–D961.

[7] B.A. Malin, K.E. Emam, C.M. O'Keefe, Biomedical data privacy: problems, perspectives, and recent advances, J. Am. Med. Inform. Assoc. 20 (1) (2013) 2–6.

[8] M.M.A. Aziz, M.N. Sadat, D. Alhadidi, S. Wang, X. Jiang, C.L. Brown, N. Mohammed, Privacy-preserving techniques of genomic data—a survey, Brief. Bioinform. 20 (3) (2019) 887–895.

[9] O. Zolotareva, R. Nasirigerdeh, J. Matschinske, R. Torkzadehmahani, M. Bakhtiari, T. Frisch, J. Späth, D.B. Blumenthal, A. Abbasinejad, P. Tieri, et al., Flimma: a federated and privacy-aware tool for differential gene expression analysis, Genome Biol. 22 (1) (2021) 1–26.

[10] S. Warnat-Herresthal, H. Schultze, K.L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers, N.A. Aziz, et al., Swarm learning for decentralized and confidential clinical machine learning, Nature 594 (7862) (2021) 265–270.

[11] D. Wiltshire, S. Alvanides, Ensuring the ethical use of big data: lessons from secure data access, Heliyon 8 (2) (2022) e08981.

[12] A. Honkela, M. Das, A. Nieminen, O. Dikmen, S. Kaski, Efficient differentially private learning improves drug sensitivity prediction, Biol. Direct 13 (1) (2018) 1–12.

[13] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference, Springer, 2006, pp. 265–284.

[14] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (3–4) (2014) 211–407.

[15] M.M. Islam, N. Mohammed, Y. Wang, P. Hu, Differential private deep learning models for analyzing breast cancer omics data, Front. Oncol. 12 (2022).

[16] J. Konečnỳ, H.B. McMahan, D. Ramage, P. Richtárik, Federated optimization: distributed machine learning for on-device intelligence, preprint, arXiv:1610.02527, 2016.

[17] J. Konečnỳ, H.B. McMahan, F.X. Yu, P. Richtárik, A.T. Suresh, D. Bacon, Federated learning: strategies for improving communication efficiency, preprint, arXiv:1610.05492, 2016.

[18] H.B. McMahan, E. Moore, D. Ramage, B.A. y Arcas, Federated learning of deep networks using model averaging, preprint, arXiv:1602.05629, 2016.

[19] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (2019) 1–19.

[20] S. Chen, D. Xue, G. Chuai, Q. Yang, Q. Liu, FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery, Bioinformatics 36 (22–23) (2021) 5492–5498.

[21] S. Sanyal, D. Wu, B. Nour, A federated filtering framework for Internet of medical things, in: ICC 2019-2019 IEEE International Conference on Communications (ICC), IEEE, 2019, pp. 1–6.

[22] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, X. Jiang, et al., Privacy-preserving patient similarity learning in a federated environment: development and analysis, JMIR Med. Inform. 6 (2) (2018) e7744.

[23] M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 92–104.

[24] I. Dayan, H.R. Roth, A. Zhong, A. Harouni, A. Gentili, A.Z. Abidin, A. Liu, A.B. Costa, B.J. Wood, C.-S. Tsai, et al., Federated learning for predicting clinical outcomes in patients with Covid-19, Nat. Med. 27 (10) (2021) 1735–1743.

[25] J.L. Salmeron, I. Arévalo, A. Ruiz-Celma, Benchmarking federated strategies in Peer-to-Peer federated learning for biomedical data, Heliyon (2023).

[26] B. Seashore-Ludlow, M. Rees, J. Cheah, M. Cokol, E. Price, M. Coletti, V. Jones, N. Bodycombe, C. Soule, J. Gould, et al., Harnessing connectivity in a large-scale small-molecule sensitivity dataset, Cancer Discov. 5 (2015) 1210–1223, https://doi.org/10.1158/2159-8290.CD-15-0235, Tech. Rep., [PMC free article][PubMed][CrossRef][Google Scholar].

[27] D. Maeser, R.F. Gruener, R.S. Huang, oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data, Brief. Bioinform. 22 (6) (2021) bbab260.

[28] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, C.J. Wilson, J. Lehár, G.V. Kryukov, D. Sonkin, et al., The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (7391) (2012) 603.

[29] B. Haibe-Kains, N. El-Hachem, N.J. Birkbak, A.C. Jin, A.H. Beck, H.J. Aerts, J. Quackenbush, Inconsistency in large pharmacogenomic studies, Nature 504 (7480) (2013) 389–393.

[30] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[32] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.

[33] J. Wang, Q. Liu, H. Liang, G. Joshi, H.V. Poor, Tackling the objective inconsistency problem in heterogeneous federated optimization, Adv. Neural Inf. Process. Syst. 33 (2020) 7611–7623.

[34] S.P. Karimireddy, S. Kale, M. Mohri, S.J. Reddi, S.U. Stich, A.T. Suresh, SCAFFOLD: Stochastic controlled averaging for on-device federated learning, 2019.

[35] T. Fawcett, An introduction to roc analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.

[36] J. Opitz, S. Burst, Macro f1 and macro f1, preprint, arXiv:1911.03347, 2019.

[37] Q. Li, Y. Diao, Q. Chen, B. He, Federated learning on non-iid data silos: an experimental study, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), IEEE, 2022, pp. 965–978.