RESEARCH ARTICLE

Check for updates

Radiomic Evaluation Model on the Efficacy of Neoadjuvant Chemotherapy for Non-small Cell Lung Cancer A Multicenter Collaborative Research Based on Privacy Protection

Yuehong Long[2], **Yuliang Zeng**[1], **Tao Zhang**[1], **Jiancun Zhou**[1], **Qian He**[1], **Xingqi Liu**[1] and **Ke Wang**[1,*]

[1] College of Information and Electronic Engineering, Hunan City University, Yiyang 413000, China
[2] School of Municipal and Geomatics Engineering, Hunan City University, Yiyang 413000, China

Abstract

Background: Practical implementation of radiomics research faces significant data accessibility challenges due to privacy and ethical restrictions on multicenter data aggregation. Federated Learning (FL) provides a secure distributed framework that preserves data privacy through cryptographic techniques. Its adoption in radiomics is an emerging trend, enabling collaborative training without sharing sensitive imaging data. However, the inherently Non-IID data distribution across clients in FL often leads to class imbalance, which can substantially degrade global model performance. Purpose: To develop a privacy-preserving, multicenter collaborative CT-radiomics model for evaluating neoadjuvant chemotherapy efficacy in non-small cell lung cancer (NSCLC). Methods: To mitigate FL performance degradation caused by data imbalance, we propose a parameter-sharing federated aggregation algorithm (FedPS), where model parameters are sequentially shared via the server. Results: On an imbalanced NSCLC NAC efficacy dataset, centralized learning achieved an AUC of 0.92. FedPS attained competitive performance (AUC = 0.88), approaching the centralized benchmark while preserving privacy. Common FL algorithms performed lower: FedAvg (AUC = 0.84), FedSGD (0.85), and FedProx (0.85). On extremely imbalanced data, FedPS maintained good performance (AUC = 0.86), compared to FedAvg (0.80), FedSGD (0.83), and FedProx (0.85). Conclusions: The proposed FedPS algorithm demonstrates promising classification and generalization performance in imbalanced federated learning scenarios.

Keywords: radiomics, federated learning, deep learning, distributed learning.

Academic Editor:
Zhunga Liu

Submitted: 24 June 2025
Accepted: 19 December 2025
Published: 04 January 2026

Vol. 3, **No.** 1, 2026.
10.62762/CJIF.2025.125241

***Corresponding authors:**
Ke Wang
wangke@hncu.edu.cn

Citation
Long, Y., Zeng, Y., Zhang, T., Zhou, J., He, Q., Liu, X., & Wang, K. (2025). Radiomic Evaluation Model on the Efficacy of Neoadjuvant Chemotherapy for Non-small Cell Lung Cancer A Multicenter Collaborative Research Based on Privacy Protection. *Chinese Journal of Information Fusion*, 3(1), 17–30.

1 Introduction

As an important branch of image analysis, radiomics reveals new features related to treatment outcome, disease molecular expression, or patient survival by extracting quantitative features from medical images (including CT, MRI, and PET) that may exceed human perception [1, 2]. On the basis of conventional CT images, CT radiomics uses deep data mining to find high-dimensional features such as shape features, morphology, boundaries, histograms, textures, and wavelet transforms, and converts these features into quantitative image features to reflect human tissue, cell and gene level changes [3, 4]. Rational use of CT radiomics can timely and accurately evaluate the effectiveness of neoadjuvant chemotherapy, and can solve the problem of "over-chemotherapy" or "under-chemotherapy" caused by the failure to adjust the treatment plan in time. This provides auxiliary decision support for the optimization of clinical treatment plans, which is of great significance for the realization of personalized tumor treatment [5, 6]. Currently, most radiomics efficacy evaluation only involve single-center and small-sample medical imaging [7], with limited generalizability and credibility. Multi-center clinical research involving multiple hospitals/medical institutions have attracted more and more attention and has become an important way for domestic and foreign hospitals or medical institutions to carry out radiomics [8–10]. However, the data of each independent medical institution presents an isolated island distribution. Moreover, constraints such as privacy security, morality and ethics make it difficult to gather image from independent institutions together for model training in real world [11, 12]. Federated learning (FL) [13, 14] is an encrypted distributed machine learning that allows each participants to achieve the purpose of jointly building a predictive model without disclosing the original data. Under the premise of protecting data privacy, the participants utilizes local data for model training, and the central server is responsible for collecting, aggregating and distributing model parameters. After multiple rounds of training, a model that is close to the result of centralized machine learning is obtained on the central server. Therefore, the application of federated learning in the field of radiomics has become an inevitable trend [15]. On the one hand, it breaks the data barriers between different hospitals due to data privacy protection, so that each participant can obtain a better model and ensure data security without sharing data. On the other hand, this multi-party collaborative model has lossless or even better performance than models based on traditional machine learning methods.

The mainstream algorithm for federated learning is the federated averaging (FedAvg) algorithm proposed in [16]. The algorithm assumes that the data on the participants is a random sampling of the overall data, while the data of each participation in federated learning is collected independently and comes from different hospitals or institutions. Therefore, there are differences between the data owned by different hospitals or institutions, resulting in the phenomenon of Non-IID [17, 18]. Among them, the most common situation is that the data volume and data label distribution are different among different hospitals or institutions. Different from the assumption that "data follows independent and identical distribution" in the derivation of common algorithms such as machine learning and deep learning, the Non-IID in federated learning directly affects the performance of the FedAvg aggregation algorithm [19].

Aiming at the problem of data Non-IID, many literatures have proposed methods to optimize the performance of federated learning. Literature [20] improves the performance of the FedAvg model by sharing 5% of the data on the CIFAR-10 dataset. Based on the FedAvg loss function, the literature [21] introduces a near-end term to limit the update range of the client and optimize the aggregation performance of the federated model in the case of Non-IID. Literature [22] allows the central server to select specific clients to participate in federated model training by using the implicit connection between model parameters and data distribution, so as to balance the deviation introduced by Non-IID. Most of the existing research focuses on MNIST, CIFAR-10 and other image classification and text prediction tasks, with the purpose of improving the performance of the federated learning algorithm on Non-IID data. However, there are few Non-IID studies on structured data. The effect of NAC in different participants in NSCLC is affected by the lack of image data, different CT model parameters and other factors, leading to the sample distribution of treatment-responsive or non-responsive was not balanced. Unlike standard image datasets such as MNIST and CIFAR-10, radiomics data is typically characterized by a small sample size, low dimensionality (following feature selection and fusion), and an imbalance in class distribution across different institutions. When the distribution of data class is unbalanced, the aggregation and performance of the federated model

Table 1. NSCLC Neoadjuvant Chemotherapy patients' information.

		Dataset 1		Dataset 2		Dataset 3	
		CR+PR (n=75)	NC+PD (n=43)	CR+PR (n=114)	NC+PD (n=63)	CR+PR (n=41)	NC+PD (n=40)
Age	Range	38-75	33-72	36-74	35-68	40-75	35-71
	Mean±SD	65.6±7.6	61.2±8.3	63.1±9.6	59.1±7.1	67.5±8.2	64.1±7.9
Sex	male	48	30	73	46	25	29
	female	27	13	41	17	16	11
Smoking	Yes	45	26	59	41	17	24
	No	30	17	55	22	24	16

are seriously affected. To address this challenge, we propose a novel federated aggregation algorithm to mitigate aggregation difficulties and performance degradation caused by data imbalance in federated models.

2 Materials and Methods

2.1 Data Acquisition

Three Grade A hospitals in Yiyang City, Changsha City, and Zhongshan City are selected to participate in the efficacy study of non-small cell neoadjuvant chemotherapy. According to the requirements of the project initiators, each hospital needs to prepare tools such as 3D-Slicer and PyRadomics, and collect NSCLC cases undergoing similar NAC between 2015 and 2021.

Inclusion criteria include:

a) The patient is between 18 and 75 years old.

b) The patient was diagnosed as NSCLC by histopathological biopsy or multidisciplinary consultation.

c) The patient has one measurable lesion.

d) The patient has a pre-treatment CT image with a scan slice thickness of 5 mm.

e) There is other demographic information about the patient, such as gender and smoking status.

f) The patient has no history of surgical resection of the tumor.

g) The patient has undergone neoadjuvant chemotherapy.

Exclusion criteria include:

a) The patient has received any anti-tumor-related treatment before admission.

b) The patient is accompanied by other primary tumors;

c) The patient voluntarily requests withdrawal without completing the treatment plan;

d) There is no curative effect evaluation information for the patient after treatment.

The neoadjuvant chemotherapy program mainly uses chemotherapy programs such as TP (paclitaxel+carboplatin), GP (gemcitabine+cisplatin) or PP (pemetrexed+ cisplatin). Paclitaxel, which prevents cancer growth by interfering with the cell cycle, and cisplatin, which destroys DNA structures, have anti-tumor effects. Both improve treatment efficacy by inhibiting tumor cells. Therefore, the above chemotherapy regimen was selected and included in the scope of this work. For the convenience of expression, the data corresponding to the three hospitals are respectively represented by dataset 1, dataset 2 and dataset 3. The specific statistics are shown in Table 1.

In Table 1, (CR+PR) represents the neoadjuvant chemotherapy sensitive group, namely good response group, while (NC+PD) represents the neoadjuvant chemotherapy insensitive group, namely poor response group.

2.2 Extraction and Selection of Radiomics Features

The 3D-Slicer is used for tumor segmentation on CT images of NSCLC, and the PyRadiomics is used to extract 875 radiomics features from ROIs of NSCLC. Considering that the subsequent radiomics analysis is for quantitative features, non-numerical features are removed. 856 quantitative features are obtained. Then, 5 informative features about medical image version, and size are removed. After preliminary processing, 851 quantitative features are finally retained. Each participating center uses a coefficient threshold-based multi-level feature selection method [23] for feature screening. Due to the difference in the final reserved feature subsets of each participation, a combination of sequence forward search (SFS) method and unsupervised evaluation (SOM) method is used to fuse multiple center feature subsets. Finally, a federated feature subset of 10 features is retained.

2.3 Method

2.3.1 FedAvg Algorithm

FedAvg is currently the most common aggregation algorithm [24], which is suitable for horizontal federation scenarios [25]. FedAvg algorithm aggregates the client model parameters in an average way, and then distributes them to the client. The specific algorithm flow is shown in Algorithm 1.

When the t round of training starts, each client receives the parameter w^{t-1} of the global model from the central server. After receiving the model parameters, each client continues to train the model using the local data. After several iterations, the local model parameter w_k is obtained. The server will receive gradient update information from the client and perform weighted average aggregation, then send the aggregated parameter information to the client. The client will utilize aggregated information for the next iteration of training.

According to the above FedAvg process, the complexity of the algorithm is controlled by three key parameters, including the proportion of clients participating in federated model training C, the client mini-batch size B, and the number of training rounds of the model on the local dataset E. C represents the proportion of clients selected by the central server. C=1 indicates that all clients participate in federated learning training. When E is too small, it increases the number of global communications. On the contrary, the federated model is prone to fall into the problem of local optimum, long aggregation time of the global model or failure to converge. This situation is exacerbated when the distribution of client data is unbalanced. The reduction of B can effectively reduce the number of communication rounds, which helps to relieve the communication pressure between the client and the central server. The experiment shows that the FedAvg can reduce the communication overhead between the client and the central server, and its performance under independent and identically distributed data is basically consistent with traditional centralized model training methods. Although FedAvg can also converge under Non-IID. Considering the unbalanced distribution of positive and negative samples between clients, the data features of the minority class will be suppressed by the data features of the majority class, resulting in divergence in the direction of model update parameters [26], that is, gradient divergence. Ultimately, problems such as slow model aggregation and low accuracy will seriously affect the performance

Algorithm 1: FedAvg

Input: Initializes model parameters w.
Output: Global model parameters w^*.

Parameters:
K : Number of clients;
C : The proportion of central server selecting clients $(0 < C \leq 1)$;
B : The local mini-batch size of the client;
E : Number of local training rounds;
η : Learning rate;

Server: ;
(1) Initializes the global model parameters w;
(2) Randomly select $m = \max\{C \times K, 1\}$ clients to participate in this federation training;
(3) The selected client performs local updates (steps can be found in the client local update section), and the server receives model parameters w_k from each client, where $k \in [1, K]$;
(4) Aggregate all client model parameters w_k, update current global model parameters $w = \sum_{k=1}^{N} n_k w^k$, and distributed to each client;
(5) Iterations (2) to (4) continuously update the global model parameters w until the stop condition is met;
(6) Training completed, return the final global model parameters w^* to all K clients;

Client Local Update (Taking the k-th client and round t as an example): ;
(1) Client k receives the current global model parameters w_k^{t-1}, as a local model initialization parameter;
(2) Partition the local dataset D_k into multiple sets of size B, represented by β;
(3) Gradient update is performed on set β, i.e. $w_k^l = w_k^{l-1} - \eta \nabla U(\beta_j; w_k^{l-1})$;
(4) Iterate over all elements in set D_k, calling (3) to continually update the model parameters w_k until the last batch;
(5) Call (4) repeatedly E times, constantly update the model parameters w_k;
(6) Return the final local model parameters w_k to the central server;

of the federated model, which is not conducive to the implementation of the federated model.

Research on the evaluation of radiomics effect of NSCLC neoadjuvant chemotherapy with Federated learning. Among them, the data comes from 3 clients (hospitals). Since Medical imaging are collected independently by each hospital, the distribution of effective patients and ineffective patients in NSCLC neoadjuvant chemotherapy in different hospitals is relatively different. In the process of aggregation of client model parameters based on Fedavg, gradient divergence may occur, which will reduce the aggregation speed and performance of the global model. In view of the imbalance of positive and negative samples between clients, it is necessary to explore a more suitable federated aggregation algorithm.

2.3.2 FL under data category imbalance

The imbalance of positive and negative samples in the dataset can lead to large differences in trained models between clients. When the FedAvg to aggregate global model parameters, there will be weight divergence, which hinders the aggregation of the prediction model under the federation framework and greatly reduces the accuracy of the model. It is assumed that two clients (k_1 and k_2) participate in federated learning training, and p_{k1} and p_{k2} represent client k_1 and k_2 distribution forms, where p_{k1}, p_{k2} obey a uniform distribution.

$$p_{k1} = p_{k2} \sim U \tag{1}$$

When the model starts training, it is assumed that the initial parameters of the model are the same, as shown in formula (2):

$$w_*^0 = w_k^0 = w_{\text{fed}}^0 \tag{2}$$

where w_*^0 is the weight for centralized learning (the 0 above w represents the initial parameter; centralized learning represents that all client data are pooled together to participate in model training), w_k^0 is the k-th client weight in federated learning, and w_{fed}^0 is the initial weight in FedAvg. The update method of w_k^0 is shown in formula (3).

$$w_*^{t+1} = w^t - \eta \nabla_{w_k^t} \sum_{i=1}^{n} L\left(f\left(x^{(i)}; \alpha, w_i^t\right), y^{(i)}\right) \tag{3}$$

where w_k^t is the centralized learning weight, t is the round, α is the model structure, L is the loss function, η is the learning rate, and $x^{(i)}, y^{(i)} \sim U$. At this point, the update process of w_k^t is shown in formula (4).

$$w_k^{t+1} = w_k^t - \eta \nabla_{w_k^t} \sum_{i=1}^{N_k} \frac{1}{N_k} \cdot L\left(f\left(x^{(i)}; \alpha, w_t^{(k)}\right), y_k^{(i)}\right) \tag{4}$$

where w_k^t is the weight trained by the client k on the local dataset, and $(x_k^{(i)}, y_k^{(i)}) \sim U$. $k \in [k_1, k_2]$, and N_k is the number of samples that client k participated in training.

The server uses the FedAvg to integrate the model parameters of multiple clients, as shown in formula (5):

$$\begin{aligned} w_{\text{Fed}}^{t+1} &= \sum_{k=1}^{K} \frac{N_k}{N} w_k^{t+1} \\ &= \sum_{k=1}^{K} \frac{N_k}{N} w_k^t \\ &- \sum_{k=1}^{K} \frac{1}{N} \cdot \eta \cdot \nabla_{\nu_k^t} \sum_{i=1}^{N_k} L\left(f\left(x_k^{(i)}; \alpha, w_k^t\right), y_k^{(i)}\right) \end{aligned} \tag{5}$$

where K and N represent the number of clients and samples participating in federated model training, respectively. That is $N = \sum_{k=1}^{K} N_k$. When $x_k^{(i)}, y_k^{(i)} \sim U$, formula (5) is further simplified.

$$w_{\text{Fed}}^{t+1} = w_k^t - \eta \nabla_{\nu_k^t} \sum_{i=1}^{N} L\left(f\left(x_k^{(i)}; \alpha, w_k^{(k)}\right), y_k^{(i)}\right) \tag{6}$$

According to formula (6) and formula (4), it can be obtained that $w_{\text{Fed}}^{t+1} \cong w_*^{t+1}$. That is, when the client data follows a uniform distribution, the centralized learning and FedAvg update model parameters are the same. With the roughly same client data distribution, FedAvg can get the same or similar parameters for centralized learning.

In the federated learning framework, in order to better describe the gradient divergence problem of FedAvg when the distribution of positive and negative samples is not uniform, it is assumed that the client k_1 obeys a uniform distribution, while k_2 does not obey a uniform distribution. In formula (5), when k is taken as k_1 and

k_2 respectively,

$$\sum_{k=0}^{\infty} L\left(f\left(x_k^{(i)}, \alpha, w_k^{(i)}\right), y_k^{(i)}\right) \neq \sum_{k=0}^{\infty} L\left(f\left(x_k^{(i)}, \alpha, w_k^{(i)}\right), y_k^{(i)}\right). \tag{7}$$

That is, there is a difference in the model parameter update between client k_1 and client k_2. When the FedAvg is used to average the two client parameters, it is obvious that

$$w_{\text{Fed}}^{t+1} \neq w_*^{t+1}. \tag{8}$$

In order to reduce the communication frequency between the client and the server, the FedAvg generally uploads the model parameters obtained through iteration to the central server after the client performs multiple rounds of local iteration. The more local iterations there are, the greater the difference in model weights between the client (k_2) and centralized learning, i.e., the greater the weight divergence. For local more iteration client k_2, formula (4) is used to calculate client-local updates. It can be concluded that:

$$w_{k_2}^{t+1} \gg w_*^{t+1} \tag{9}$$

When aggregating parameters by using formula (5), the federated parameters w_{Fed}^{t+1} are updated as shown in (8):

$$w_{\text{Fed}}^{t+1} = \frac{N_k}{N} w_{k_1}^{t+1} + \frac{N_{k_2}}{N} w_{k_2}^{t+1} \tag{1.8}$$

When the client k_1 obeys the uniform distribution, $w_{k_1}^{t+1} \cong w_*^{t+1}$. In formula (8), when the positive and negative samples are unbalanced, the more the federated model parameter w_{Fed}^{t+1} deviates from the centralized model parameter w_*^{t+1}. That is to say, when the distribution of data categories is unbalanced, the FedAvg cannot achieve the optimal weight. As the imbalance of data categories increases, the weight divergence will become more significant, making it difficult for the model to aggregate on the central server.

In theory, it is obviously impossible to use the FedAvg to aggregate client model parameters and obtain global optimal model parameters for multiple low-correlation client model parameters. Based on the FedAvg, this paper will discuss the federated model aggregation method to reduce the negative impact of the unbalance of client data categories, so as to improve the predictive performance of the federated model.

Algorithm 2: Parameter Sharing Federated Learning Algorithm

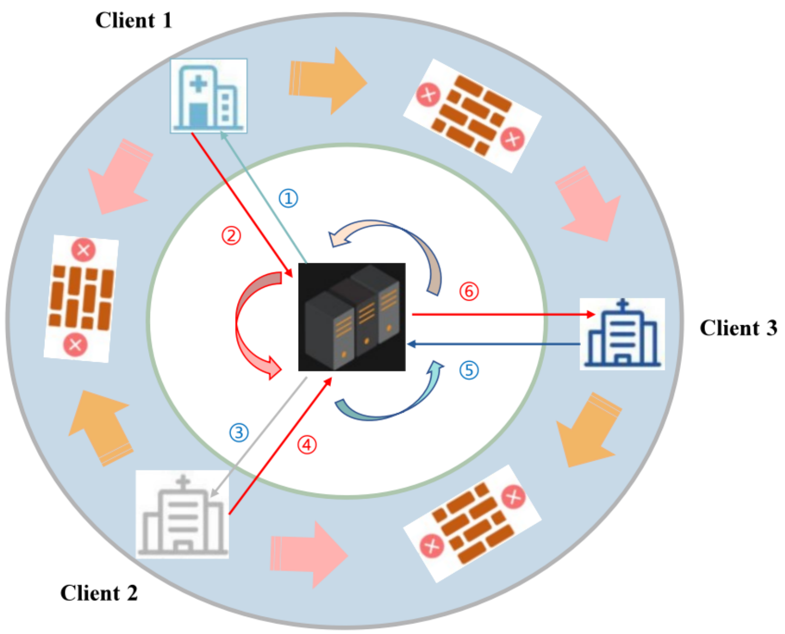
Input: Initialize model parameter w_0.
Output: Global model parameter w^*.

Parameters:
K : number of clients;
D_i : the training data in the i-th client;

(1) Model initialization: the central server initializes the model parameters w_0;
(2) Weight sharing: ;
 a) When $i \in [1, K-1]$, the i-th client downloads the model parameter w_{i-1} from the central server, and uses the local data D_i to perform a round of model training. The obtained model parameter w_i is uploaded to the central server;
 b) When $i = K$, the last client downloads the model parameter w_{i-1} from the central server, and uses the local data D_i to perform a round of model training. The obtained model parameter w_i is uploaded to the central server, and setting $w_0 = w_i$. A round of parameter sharing is completed;
(3) Set the number of rounds: Repeat step (2) until the set number of rounds is met;

There has always been a problem with catastrophic forgetting in the field of deep learning. Catastrophic forgetting refers to the possibility of forgetting the knowledge learned on old datasets when training models on new datasets. The task of using a deep learning model to identify pictures containing cats and dogs is taken as an example. Dataset a and dataset b are composed of individual cat and dog pictures, respectively. First, the data of client a is used to train the deep learning model M. Model M is able to capture the features of cats and accurately recognize pictures of cats. Then, the model M continues training on the client b which only contains dog pictures. At this point, the model M can obtain the features of the dog and accurately identify the picture of the dog. However, the model M cannot accurately recognize cat pictures because it has forgotten the cat features.

Performing model training on dataset a and dataset b respectively is equivalent to completing two different tasks. When the model learns new knowledge, the old knowledge will be replaced by new knowledge.



Where ① ③ ⑤ represents downloading parameters from the server;
② ④ ⑥ represents uploading parameters to the server

Figure 1. The parameter passing process in one round of FedPS.

Alternative learning of new and old knowledge refers to learning old task knowledge while learning new tasks, which can avoid catastrophic forgetting. Therefore, the design of the federated learning algorithm can adopt a method similar to alternate learning to avoid the catastrophic forgetting in deep learning.

Therefore, the design of federated aggregation algorithm for unbalanced categories datasets needs to pay attention to the following two issues:

1. FedAvg aggregation model parameters are used. Datasets with unbalanced distribution of positive and negative samples are prone to gradient divergence phenomenon. The more local iterations the client model has, the more gradient divergence it will exacerbate.

2. For datasets with unbalanced distribution of positive and negative samples, catastrophic forgetting exists in the training of deep learning models.

To solve the above problem, the number of local iteration needs to be reduced to avoid gradient divergence. Furthermore, a method similar to alternating learning can solve catastrophic forgetting. In response to this, the parameter sharing federated learning aggregation algorithm (FedPS) for unbalanced radiomics data was proposed, referred to as the parameter sharing federated learning algorithm. See Algorithm 2 for details:

Algorithm 2 mainly solves catastrophic forgetting and gradient divergence in the model training process through the following two aspects.

1. The number of client iterations is set to 1.

2. The client model parameters are uploaded to the central server. The next client downloads the model parameters of the previous client from the central server. The training of the local model is performed so that the federated model is alternately trained on different clients.

In the model training process of Algorithm 2, the model parameters are shared between adjacent clients through the central server, which is similar to the centralized model training process. The client model uses local data D_i for training to avoid catastrophic forgetting and gradient divergence. Figure 1 describes the parameter passing process in one round of FedPS.

Figure 1 completes the parameter sharing process in the order of ①-②-③-④-⑤-⑥, which is a round of training. For the convenience of expression, NSCLC curative effect evaluation federal database 1 is represented by client 1, and so on. Client 1 starts

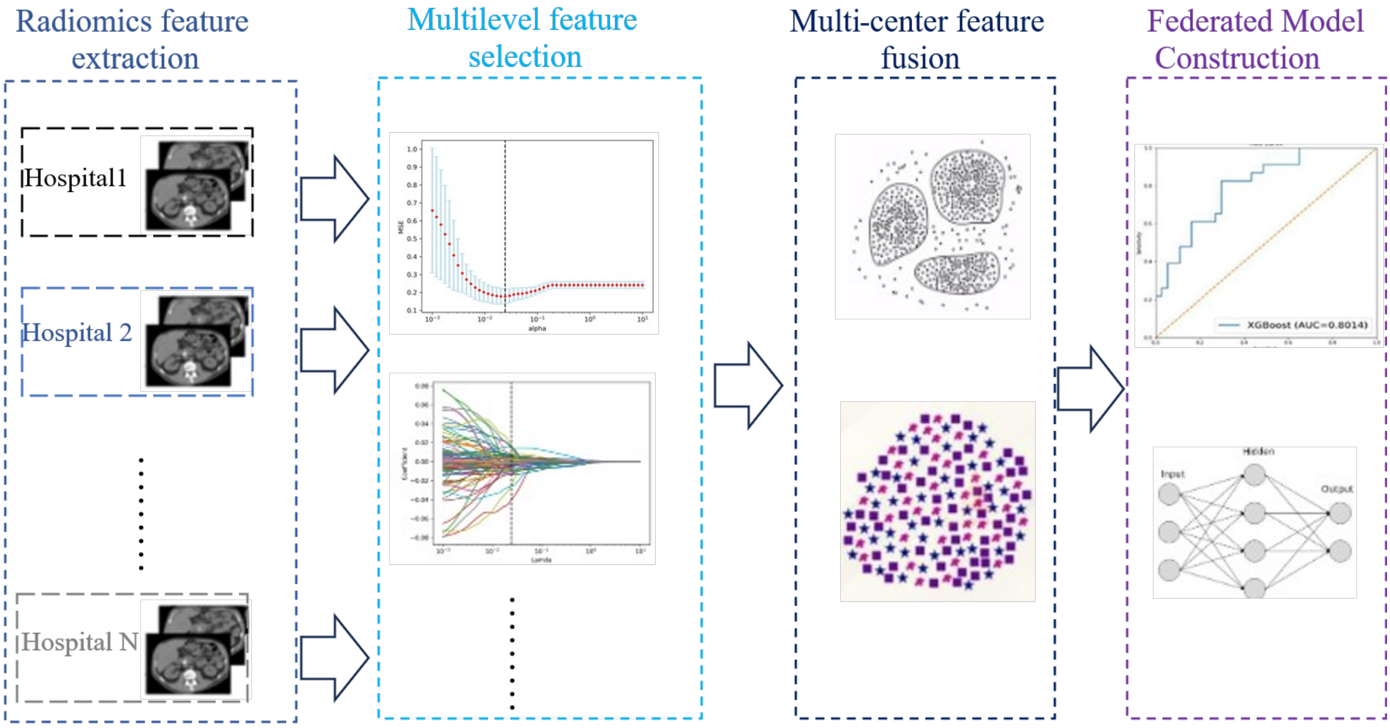


Figure 2. The parameter passing process in one round of FedPS.

training the model using the parameters of client 3 in the previous round, and client 2 uses the model parameters of client 1 to train the model, and so on. This is similar to training models in batches in centralized machine learning.

Compared with FedAvg, this algorithm needs to frequently pass and share parameters with the central server, resulting in relatively high communication costs. In this paper, research on the curative effect evaluation of radiomics based on federated learning was carried out. Participating clients, data samples, and data passed between clients and central server parameters are few. In this case, the impact of communication cost on the algorithm is almost negligible.

The workflow of introducing federated learning into the process of radiomics model construction is shown in Figure 2.

3 Result

This study utilizes a federated database for evaluating the efficacy of neoadjuvant chemotherapy in NSCLC. To address the issue of class imbalance within client data, a parameter-sharing federated learning algorithm is proposed. The data distribution for the three clients involved in the experiment is presented in Table 1.

All three clients participate in each round of federated learning training, that is, C=1. Each client builds the same deep learning network and trains the model once on the local data. Mini-batch gradient descent is chosen for updating the local model parameters. In the central server, the parameter sharing is used to transfer model parameters. The deep learning network structure of the client is shown in Figure 3.

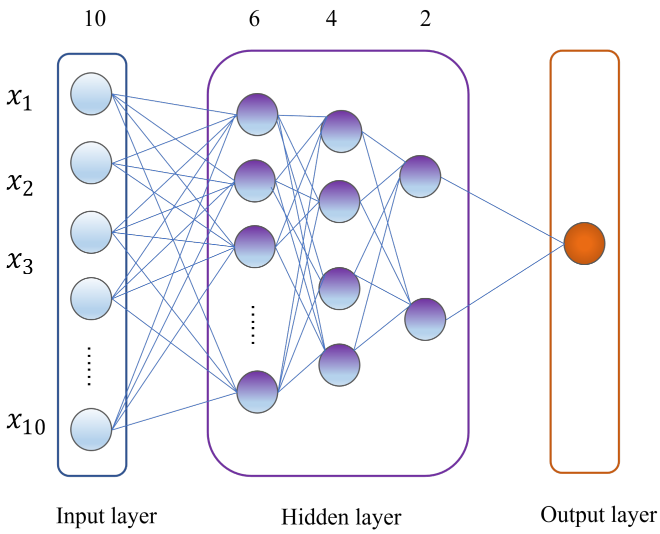


Figure 3. Neural network structure.

Deep neural network consists of an input layer, a hidden layer, and an output layer. Among them, the input layer consists of 10 neurons, which is the number of features retained after feature selection. The hidden layer consists of 3 layers, in which the number of neurons is 6, 4 and 2 in sequence. In addition, hidden layer neurons have a built-in ReLu function for nonlinear transformation. The output layer consists of one neuron. Through the Sigmoid function, the input of the previous layer is converted into a probability value output. Other hyperparameters are set to Batch-Size=8, and Learning Rate = 0.01. It was divided into training set and test set according to the ratio of 8:2.

We intentionally employed a simple deep neural network architecture for this study. The primary objective of our work was to demonstrate and validate the core mechanism of the FedPS algorithm—specifically, its capability to mitigate the challenges of data imbalance in a federated learning setting. To achieve this proof-of-concept, we utilized a standard, lightweight architecture (e.g., a simple multilayer perceptron), which is commonly adopted in foundational federated learning research. This deliberate choice was made to minimize confounding factors and to clearly attribute any performance improvements to our aggregation strategy, rather than to model complexity.

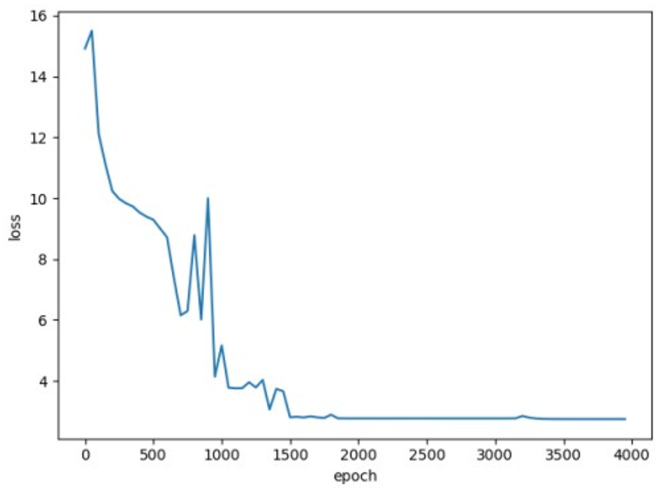


Figure 4. FedPS training process.

We chose the PySyft framework to conduct federated model training. Figures 4 and 5 show that the model was trained for 4000 rounds with the above parameters and dataset, and the results show that after about 1800 rounds of training, the model tends to stabilize. The final AUC on the test set was 0.88. In order to compare the performance of FedPS, we chooses FedAvg, FedSGD, FedProx, centralized learning and other methods for comparison. For a clearer comparison, Table 2 presents a comparative summary of the differences between FedPS and FedAvg, FedSGD, and FedProx.

Table 2. Comparison with baseline federated algorithms.

	Local Iteration Count	Parameter averaging	Limit update magnitude
FedProx	multiple	Yes	Yes
FedAvg	multiple	Yes	No
FedSGD	one	Yes	No
FedPS	one	No	No

It can be observed from Table 2 that a primary divergence lies in the local iteration count: FedPS and FedSGD are limited to one iteration, in contrast to the multiple iterations of FedAvg and FedProx. Another key difference is the aggregation process; FedPS is distinct in operating without central parameter averaging, a requirement for the other method. Finally, FedProx is uniquely characterized by the imposition of a constraint on local update magnitudes, a feature absent in the compared approaches.

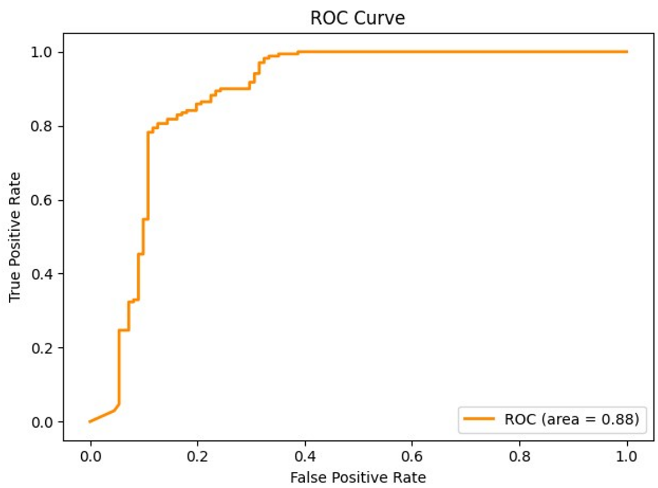


Figure 5. Classification performance of FedPS.

The methods mentioned above all adopt the same network model and hyperparameters as FedPS. Table 3 shows the classifier performance under different methods.

Table 3 shows the mean and standard deviation of AUC, ACC, Recall and F1 under 20 random repetitions, expressed in the form of mean $\pm$ standard deviation (std). When the model is trained with centralized learning, the prediction performance achieved is significantly better than other methods. When FedAvg, FedSGD, FedProx, FedPS and other federated aggregation algorithms are used to train the model, the proposed FedPS is closer to centralized learning in performance. In order to compare the stability of the models, each performance index is set to standard deviation. In addition, their standard deviations on different methods were calculated separately. The results show that FedPS algorithm obtains the minimum standard deviation except for centralized learning, which directly indicates that FedPS algorithm has good stability.

To comprehensively assess the predictive reliability of the model, we employed a calibration curve (Figure 6) to evaluate the agreement between predicted probabilities and actual observations. The curve visually reflects the calibration performance of the model on a probability scale. In general, the calibration curve of a well-performing predictor approximately follows the reference line y = x (red line), with minor deviations expected, rather than overlapping perfectly with it. The degree of agreement between the calibration curve (blue line in the figure) and the reference line serves as an important metric for evaluating the model: the closer the calibration curve is to the reference line, the more reliable the model's predictions are considered to be; conversely, greater deviation indicates poorer predictive performance. Moreover, if the calibration curve lies entirely above the reference line, it suggests that the model systematically underestimates the probability of a positive outcome; if it lies below, the model tends to overestimate the probability of a positive outcome.

A key finding is that FedPS and FedSGD achieve the best calibration performance among the tested algorithms. Their calibration curves align most closely with the reference line (y=x), which translates into more reliable predictions characterized by greater accuracy and lower uncertainty in the probability estimates.

To verify the generalization performance of the proposed FedPS algorithm, we constructed three highly unbalanced datasets by randomly drawing samples without replacement from the three original datasets. A classifier is known to be biased towards the majority class in binary classification when the imbalance ratio (IR) exceeds 4:1. To investigate this under extreme conditions, our experimental setup assigned an IR greater than 4:1 to each of the three local hospitals, whereas the overall sample IR was maintained below this threshold. The detailed data distributions are provided in Table 4.

To simulate real-world data distributions and further investigate the performance of FedPS on extremely imbalanced datasets, we repeated the previous section's experiment on such datasets. Accordingly, Table 5 summarizes the classification performance of the different methods under these extreme conditions.

Table 5 shows the performance of the model in cases where the dataset is extremely imbalanced. Among them, the proposed FedPS achieves similar results on imbalanced datasets.

(1) On extremely unbalanced datasets, the performance of the four federated aggregation algorithms all decline to varying degrees. For example, the AUC, ACC, Recall, and F1 of the FedAvg on extremely unbalanced datasets are 0.8, 0.78, 0.75, and 0.78, respectively, which are 4.8%, 4.9%, 10.7% and 4.9% lower than the performance on unbalanced data. The corresponding FedPS algorithm is only down 2.8%, 4.5%, 6.8% and 4.5%.

(2) Based on the FedAvg, FedSGD sets the number of local iterations to 1 time. The imbalance between positive and negative samples can alleviate the gradient divergence in federated aggregation. In Table 3 and Table 5, the AUC of the FedSGD aggregation is higher than that of the FedAvg, which verifies that the reduction of the number of local iterations on the unbalanced dataset is beneficial to improve the performance of the federated aggregation algorithm.

(3) In Table 3 and Table 5, the standard deviation of DNN model prediction performance (such as AUC and ACC) based on FedPS aggregation algorithm is relatively low, and it is better than other federated aggregation algorithms, which shows that the DNN model based on the FedPS aggregation algorithm has good stability.

In summary, it can be concluded that the performance of the proposed FedPS is better than other federated aggregation algorithms.

4 Discussion

Radiomics uses computer technology to extract a large number of features that cannot be recognized by human eyes from medical images to describe tumor heterogeneity, which plays an important role in tumor screening, clinical diagnosis, treatment options selection, and curative effect evaluation. According to existing research, radiomics modeling based on

Table 3. Classifier performance on unbalanced datasets.

Aggregation algorithm	AUC	ACC	Recall	F1	Pre	Spec
FedAvg	0.84±0.037	0.82±0.021	0.83±0.022	0.82±0.021	0.81±0.024	0.81±0.026
FedSGD	0.85±0.049	0.82±0.032	0.83±0.023	0.82±0.032	0.81±0.028	0.81±0.030
FedProx	0.85±0.038	0.83±0.022	0.84±0.029	0.82±0.024	0.80±0.019	0.82±0.018
CL	0.92±0.063	0.92±0.144	0.90±0.190	0.92±0.145	0.94±0.180	0.94±0.016
FedPS	**0.88±0.015**	**0.89±0.018**	**0.88±0.021**	**0.89±0.019**	**0.90±0.018**	**0.90±0.020**

Note: CL stands for centralized learning. Recall, Spec, and Pre are abbreviations for sensitivity, specificity, and precision, respectively.

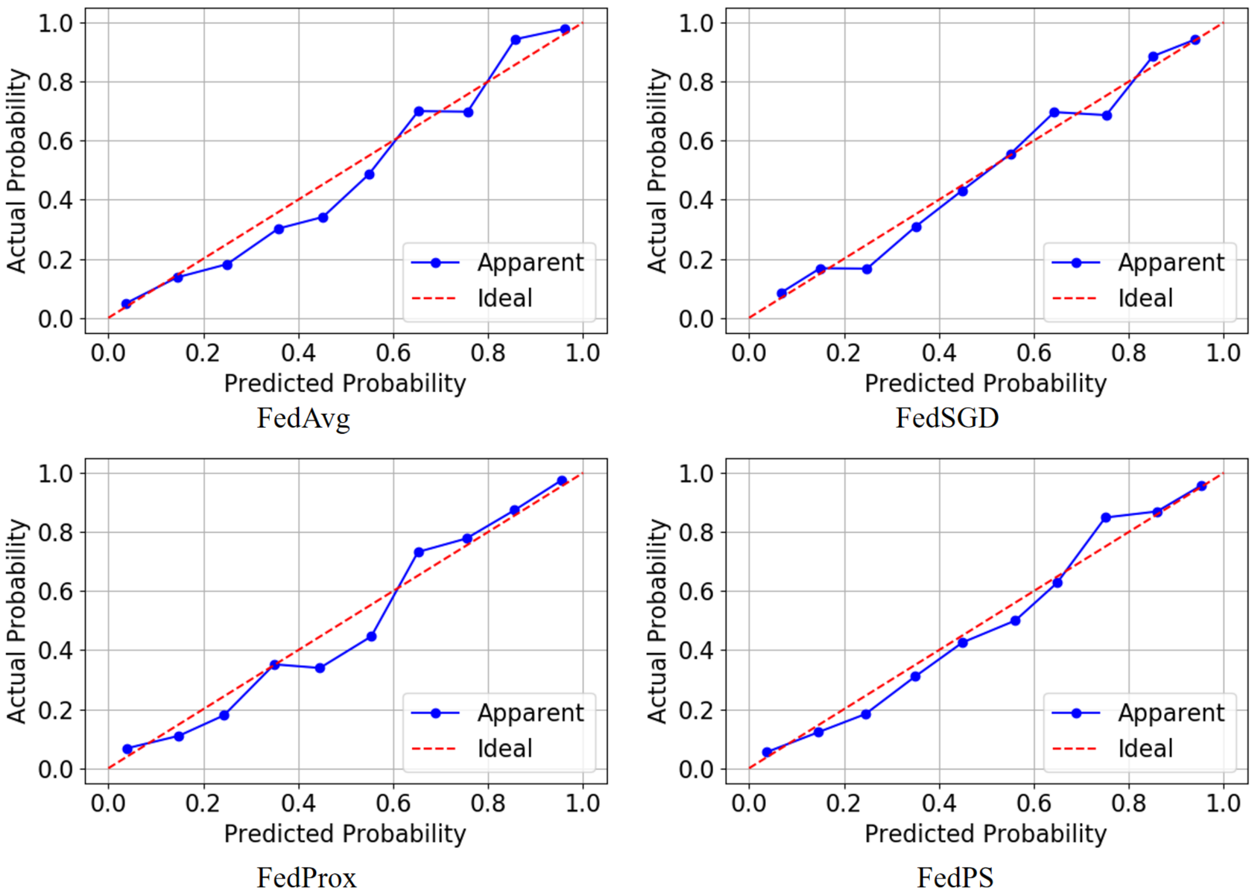


Figure 6. Calibration curves under different federal aggregation algorithms.

Table 4. Extremely unbalanced datasets.

	Dataset1	Dataset2	Dataset3
CR+PR	13	63	40
NC+PD	60	14	9
Unbalance ratio	>4:1	>4:1	>4:1

Table 5. Classifier performance on unbalanced datasets.

	AUC	ACC	Recall	F1
FedAvg	0.80±0.098	0.78±0.071	0.75±0.086	0.78±0.071
FedSGD	0.83±0.057	0.74±0.077	0.74±0.078	0.75±0.078
FedProx	0.85±0.251	0.80±0.133	0.79±0.218	0.80±0.133
Centralized learning	0.92±0.160	0.88±0.047	0.86±0.224	0.88±0.197
FedPS	**0.86±0.050**	**0.85±0.060**	**0.82±0.050**	**0.85±0.060**

artificial intelligence algorithms has an accuracy rate reaching the level of clinicians. However, traditional machine learning and deep learning algorithms need to collect a large amount of medical imaging data for model training, which violates the principle of non-sharing of private data. Therefore, this paper applies federated learning technology to the study of the efficacy evaluation of NSCLC neoadjuvant chemotherapy radiomics. On the premise of not sharing local data, the value of data scattered in various medical institutions is aggregated and the model is jointly trained. After multiple rounds of iteration and training, the NSCLC neoadjuvant chemotherapy efficacy evaluation model was formed, which can accurately predict the efficacy of NSCLC neoadjuvant chemotherapy (AUC=0.88). In the application scenario of federated learning, the data is collected independently by the participants. Each dataset basically does not meet independent and identical distribution. Among them, client data category imbalance is common. Currently, common federated learning algorithms such as FedAvg, FedSGD, and FedProx typically use aggregation and updating of client model parameters, and their performance on imbalanced datasets needs further improvement. In order to improve the performance of the algorithm, the FedPS was proposed. On the one hand, it can avoid catastrophic forgetting and gradient divergence in deep learning. On the other hand, its training process is close to the centralized machine learning training process, which can reduce the performance difference caused by the aggregation of model parameters.

(1) In order to verify the effectiveness of the proposed parameter sharing algorithm (FedPS) on the NSCLC neoadjuvant chemotherapy efficacy evaluation dataset, FedPS was compared with FedAvg, FedSGD, FedProx, and centralized learning algorithms. On the NSCLC neoadjuvant chemotherapy efficacy evaluation dataset, through the comparison of performance index such as AUC, ACC, Recall, and F1, it can be found that the performance of the FedPS is very close to that of traditional centralized learning. On the extremely unbalanced data, the FedPS also achieves similar results, which further verifies its effectiveness. The FedPS uses the parameter transfer method to share model parameters and a similar serial method to train the federated model, which increases the training time of the federated model. Due to the small number of clients participating in the NSCLC neoadjuvant chemotherapy efficacy evaluation and the small data sample size, clients training deep learning models on high-performance equipment can reduce the waiting time between clients. In this scenario, the impact of time on the algorithm is almost negligible.

(2) The unbalanced distribution of positive and negative samples between participation is a common scenario in federated learning. Due to differences in regions, grades, and data integrity among different hospitals, there is often an imbalance in the collected radiomics data for NSCLC efficacy evaluation. In order to verify the performance of the FedPS algorithm in this case, FedPS is compared with models such as FedAvg, FedSGD, FedProx, and Centralized Learning. On the extremely unbalanced NSCLC efficiency evaluation dataset, the performance of the FedPS is close to that of traditional centralized learning, which verifies its generalization performance.

(3) The proposed algorithm has certain limitations. First, the expansion of clients or samples participating in federated learning will increase the transmission and time costs of training models. Secondly, when the client fails to train the local model due to unexpected reasons such as power failure or system crash, the federated model training will be automatically terminated. These factors will affect the training of the federated model.

5 Conclusion

In this paper, we propose a federated aggregation algorithm based on parameter sharing to address the scenario of imbalanced positive and negative samples across different participants. The algorithm achieves strong classification performance on a non-small cell lung cancer neoadjuvant chemotherapy efficacy evaluation dataset (AUC = 0.86). Furthermore, it demonstrates robust classification results on extremely imbalanced datasets, confirming its generalization capability.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported in part by the National Science Foundation of Hunan Province under Grant 2023JJ50354 and Grant 2024JJ7078; in part by the Scientific research project of Hunan Education Department under Grant 24A0575.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

This multicenter retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study protocol was approved by the ethics committees of the participating institutions. The requirement for informed consent was waived by the ethics committees due to the retrospective nature of the study and the use of anonymized data.

References

[1] Doran, S. J., Kumar, S., Orton, M., d'Arcy, J., Kwaks, F., O'Flynn, E., ... & Koh, D. M. (2021). "Real-world" radiomics from multi-vendor MRI: an original retrospective study on the prediction of nodal status and disease survival in breast cancer, as an exemplar to promote discussion of the wider issues. *Cancer Imaging, 21*(1), 37. [CrossRef]

[2] Wang, K., An, Y., Li, N., Zhou, J., & Chen, X. (2022). Robust Identification of Subtypes in Non-Small Cell Lung Cancer Using Radiomics. *Traitement du Signal, 39*(4). [CrossRef]

[3] Wang, J., Wang, S., Yu, H., Liu, C., Zhao, X., Sun, S., ... & Wu, X. (2019). JCSE01. 18 A CT-Based Radiomics Approach to Predict PD1 Inhibitor Response in Non-Small-Cell Lung Cancer. *Journal of Thoracic Oncology, 14*(10), S131-S132. [CrossRef]

[4] Liu, C., Gong, J., Yu, H., Liu, Q., Wang, S., & Wang, J. (2021). A CT-based radiomics approach to predict nivolumab response in advanced non-small-cell lung cancer. *Frontiers in Oncology, 11*, 544339. [CrossRef]

[5] Lee, G., Bak, S. H., Lee, H. Y., Choi, J. Y., Park, H., Lee, S. H., ... & Lee, K. S. (2019). Measurement variability in treatment response determination for non–small cell lung cancer: improvements using radiomics. *Journal of Thoracic Imaging, 34*(2), 103-115. [CrossRef]

[6] Saw, S. P., Ong, B. H., Chua, K. L., Takano, A., & Tan, D. S. (2021). Revisiting neoadjuvant therapy in non-small-cell lung cancer. *The Lancet Oncology, 22*(11), e501-e516. [CrossRef]

[7] Cousin, F., Louis, T., Dheur, S., Aboubakar, F., Ghaye, B., Occhipinti, M., ... & Hustinx, R. (2023). Radiomics and delta-radiomics signatures to predict response and survival in patients with non-small-cell lung cancer treated with immune checkpoint inhibitors. *Cancers, 15*(7), 1968. [CrossRef]

[8] Yang, F., Chen, W., Wei, H., Zhang, X., Yuan, S., Qiao, X., & Chen, Y. W. (2021). Machine learning for histologic subtype classification of non-small cell lung cancer: a retrospective multicenter radiomics study. *Frontiers in Oncology, 10*, 608598. [CrossRef]

[9] Orlhac, F., Frouin, F., Nioche, C., Ayache, N., & Buvat, I. (2019). Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology, 291*(1), 53-59. [CrossRef]

[10] Cui, Y., Zhang, J., Li, Z., Wei, K., Lei, Y., Ren, J., ... & Gao, X. (2022). A CT-based deep learning radiomics nomogram for predicting the response to neoadjuvant chemotherapy in patients with locally advanced gastric cancer: a multicenter cohort study. *EClinicalMedicine, 46*. [CrossRef]

[11] Linardos, A., Kushibar, K., Walsh, S., Gkontra, P., & Lekadir, K. (2022). Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports, 12*(1), 3551. [CrossRef]

[12] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports, 10*(1), 12598. [CrossRef]

[13] Narmadha, K., & Varalakshmi, P. (2022). Federated learning in healthcare: a privacy preserving approach. In *Challenges of Trustable AI and Added-Value on Health* (pp. 194-198). IOS Press. [CrossRef]

[14] Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine, 158*, 106848. [CrossRef]

[15] Ali, H., Alam, T., Househ, M., & Shah, Z. (2022). Federated learning and internet of medical things–opportunities and challenges. *Advances in Informatics, Management and Technology in Healthcare*, 201-204. [CrossRef]

[16] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

[17] Chai, B., Liu, K., & Yang, R. (2022). Cross-Domain Federated Data Modeling on Non-IID Data. *Computational Intelligence and Neuroscience, 2022*(1), 9739874. [CrossRef]

[18] Traverso, A., Wee, L., Dekker, A., & Gillies, R. (2018). Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics, 102*(4), 1143-1158. [CrossRef]

[19] Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

[20] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

[21] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems, 2*, 429-450.

[22] Wang, H., Kaplan, Z., Niu, D., & Li, B. (2020, July). Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE conference on computer communications* (pp. 1698-1707). IEEE. [CrossRef]

[23] Wang, K., An, Y., Zhou, J., Long, Y., & Chen, X. (2023). A novel multi-level feature selection method for radiomics. *Alexandria Engineering Journal, 66*, 993-999. [CrossRef]

[24] Lee, J., Sun, J., Wang, F., Wang, S., Jun, C. H., & Jiang, X. (2018). Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR medical informatics, 6*(2), e7744. [CrossRef]

[25] Dhada, M., Jain, A. K., & Parlikad, A. K. (2020). Empirical convergence analysis of federated averaging for failure prognosis. *IFAC-PapersOnLine*, *53*(3), 360-365. [CrossRef]

[26] Subramanian, M., Rajasekar, V., VE, S., Shanmugavadivel, K., & Nandhini, P. S. (2022). Effectiveness of decentralized federated learning algorithms in healthcare: a case study on cancer classification. *Electronics*, *11*(24), 4117. [CrossRef]

Yuehong Long received the Ph.D. degree in Geographic Information System from Hunan Normal University, ChangSha, China. She is currently an lecturer with the College of Municipal and Surveying Engineering College, Hunan City University, YiYang, China. He has authored more than 10 peer reviewed publications. His current research interest include GIS, Privacy Computing and Medical image processing. (Email: yhlong@hncu.edu.cn)

Yuliang Zeng received the B.S. in Surveying and Mapping Engineering from Hunan City University, China, and is currently pursuing the M.S. in Electronic Information at the School of Information and Electronic Engineering, Hunan City University, China. His research interests include medical image processing and artificial intelligence. (Email: 1464942170@qq.com)

Tao Zhang received the B.S. in Data Science and Big Data Technology from Hubei University of Science and Technology, China, and is currently pursuing the M.S. in Electronic Information at the School of Information and Electronic Engineering, Hunan City University, China.His research interests include healthcare big data and artificial intelligence.

Jiancun Zhou received the Ph.D. degree in computer science and engineering, Central South University, Changsha, China. He is currently an Professor with the College of Information Science and Electronic Engineering, Hunan City University, YiYang, China. His research interests include medical big data, medical image processing and artificial intelligence for analyzing medical data.

Qian He received the Ph.D. degree in Applied Mathematics from Hunan University, ChangSha, China. He is currently an Professor with the College of Information Science and Electronic Engineering, Hunan City University, YiYang, China. He has authored more than 10 peer reviewed publications. His current research interest include smart healthcare, medical image reconstruction, He is a regular reviewer of many reputed journals and conferences.

Xinqi Liu received the master's degree in computer science and technology from Beijing University of Posts and Telecommunications, Beijing, China, in 2021. He is currently an assistant professor with Hunan City University, Yiyang, China. His research interests include medical big data and artificial intelligence for medical applications.

Ke Wang received the Ph.D. degree in medical information management from Central South University, ChangSha, China. He is currently a Associate Professor with the College of Information Science and Electronic Engineering, Hunan City University, YiYang, China. His current research interests include medical image processing, information fusion and Privacy Computing. (Email: wangke@hncu.edu.cn)