Federated Active Learning for Multicenter Collaborative Disease Diagnosis

Xing Wu*, Jie Pei, Cheng Chen, Yimin Zhu, Jianjia Wang, Quan Qian, Jian Zhang, Qun Sun, Yike Guo

***Abstract*— Current computer-aided diagnosis system with deep learning method plays an important role in the field of medical imaging. The collaborative diagnosis of diseases by multiple medical institutions has become a popular trend. However, large scale annotations put heavy burdens on medical experts. Furthermore, the centralized learning system has defects in privacy protection and model generalization. To meet these challenges, we propose two federated active learning methods for multicenter collaborative diagnosis of diseases: the Labeling Efficient Federated Active Learning (LEFAL) and the Training Efficient Federated Active Learning (TEFAL). The proposed LEFAL applies a task-agnostic hybrid sampling strategy considering data uncertainty and diversity simultaneously to improve data efficiency. The proposed TEFAL evaluates the client informativeness with a discriminator to improve client efficiency. On the Hyper-Kvasir dataset for gastrointestinal disease diagnosis, with only 65% of labeled data, the LEFAL achieves 95% performance on the segmentation task with whole labeled data. Moreover, on the CC-CCII dataset for COVID-19 diagnosis, with only 50 iterations, the accuracy and F1-score of TEFAL are 0.90 and 0.95, respectively on the classification task. Extensive experimental results demonstrate that the proposed federated active learning methods outperform state-of-the-art methods on segmentation and classification tasks for multicenter collaborative disease diagnosis.**

***Index Terms*— Federated learning, active learning, multicenter, labeling-efficient, training-efficient**

I. Introduction

With the wide use of deep learning in the field of medical imaging, researchers have been developing Artificial Intelligence (AI) algorithms [1]–[3] to improve the diagnostic accuracy of the disease. Although deep learning shows great performance in recent reports [4], [5], some worrying problems are still blocking the large-scale application of AI technologies in clinical diagnosis. Firstly, current centralized deep learning methods are not providing any protection for patients' privacy. Secondly, the trained models often fail to datasets which have different data distribution from the training dataset. What's more, current training pipeline is not suitable for an expected decentralized collaboration against the fast spread pandemics or other diseases. Researchers and medical experts are eager for a better collaboration mode in disease diagnosis.

Recent reported collaborative learning methods enable multicenter model training without sharing patient data. Among these collaborative learning methods, Federated Learning (FL) leads to privacy protection where multiple collaborators train local models on their own dataset in parallel and then upload model parameters to a central server, where local parameters are updated to global model and this model is transmitted to all participating clients for further training or application. Federated learning methods provide great promise to connect fragmented data sources as well as preserving privacy.

Although data privacy problem can be partly addressed with federated learning, this approach leads to low efficiency for collaborative diagnosis. For example, with the rapid spread of COVID-19. The time consumption for learning a diagnostic model with CT scans is mainly attributed to two aspects: largescale CT data labeling and global model training [6]. The efficiency in these two aspects is likely to be improved with Active Learning methods.

The active learning is a popular and emerging research field because the quantity and quality of data is very important for deep neural networks. The main idea of active learning is to select the most informative data samples according to specific sampling strategies. In the pool-based active learning methods [7]–[10], informative data samples are selected from an unlabeled data pool according to data uncertainty or data diversity. Then, an oracle [11] is requested to provide ground-truth labels for the selected data samples. The human-in-the-loop active learning leads to striking improvements in labeling efficiency for deep learning applications. Besides reducing labeling efficiency, we propose that the informativeness of clients in federated learning can also be evaluated and selected to boost training efficiency and reduce transmission cost.

To reduce labelling costs and protect patients' privacy, we propose federated active learning methods for multicenter

This work is supported by the National Natural Science Foundation of China (Grant No. 62172267), the National Key R&D Program of China (Grant No. 2019YFE0190500), the Natural Science Foundation of Shanghai, China (Grant No. 20ZR1420400), the State Key Program of National Natural Science Foundation of China (Grant No. 61936001), the Key Research Project of Zhejiang Laboratory (No. 2021PE0AC02).

X. Wu is with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. He is also with the Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China.(e-mail: xingwu@shu.edu.cn)

J. Pei, C. Chen, Y. Zhu, J. Wang, and Q. Qian are with the School of Computer Engineering and Science, Shanghai University, Shanghai, China.

J. Zhang is with the Shanghai Universal Medical Imaging Diagnostic Center, Shanghai, China.

Q. Sun is with Shanghai Sixth People's Hospital Affiliated to Shanghai JiaoTong University, Shanghai, China.

Y. Guo is with Hong Kong Baptist University, Kowloon Toon, Hong Kong, China.

collaborative diagnosis [12]. We propose Labeling-Efficient Federated Active Learning (LEFAL) and Training-Efficient Federated Active Learning (TEFAL) to improve labeling and training efficiency respectively. On the one hand, we propose to integrate active learning loops into each client to select informative data samples and request radiology experts to provide ground truth labels in the proposed LEFAL architecture. On the other hand, we design a discriminator to evaluate the informativeness of all clients to select the most informative ones in each federated round in the proposed TEFAL architecture. In addition, we evaluated the effectiveness and efficiency of the proposed Lefal and Tefal in segmentation and classification tasks on the CC-CCII [13]dataset and the Hyper-kvasir [14] dataset. The experimental results demonstrate that the proposed federated active learning method outperforms state-of-the-art methods in labeling and training efficiency on both segmentation and classification tasks under the multicenter collaborative settings.

The main contributions of this paper can be summarized as follows:

- We propose two federated active learning methods for multicenter collaborative disease diagnosis with high efficiency and privacy protection.
- We propose a task-agnostic hybrid sampling strategy to blend data uncertainty and data diversity to improve labeling efficiency in LEFAL, in which the concept of dataset informativeness is proposed to dynamically adjust the aggregation weights of local models.
- We design the evaluation metrics of the client informativeness and train a discriminator to select informative clients for high training efficiency in TEFAL.

II. RELATED WORK

Computer aided diagnosis system has been widely used in medical field. For example, the collection, clarity, reconstruction, processing and analysis of CT, MRI, PET and other data of patients can not be separated from the assistance of computers. With the development of deep learning techniques, federated learning are gradually introduced into the medical field in order to perform safe and efficient collaborative diagnosis between different medical institutions. Furthermore, as deep learning techniques rely on a large and even distribution of accurately annotated data points, and while more medical datasets are becoming available, the time, cost and effort required to annotate such datasets remains significant, and there is also an increasing shortage of experienced doctors to interpret medical datasets, indicating a clear need for reliable automated methods to alleviate the growing burden on healthcare practitioners.

A. Federated Learning

With the continuous development of deep learning technology in the medical imaging field, the size of deep learning models for semi-automatic or automatic diagnosis is increasing as well as the size of medical datasets is also growing, both of which make the computational overhead increase dramatically [15], [16]. Moreover, due to the huge storage cost of medical datasets, it is difficult to transfer data between datasets from different hospitals [17]. Therefore, to improve the efficiency of multicenter federated learning methods are proposed to train deep learning models for collaborative diagnosis among various hospitals [18]. Multicenter federated learning refers to the collaborative diagnosis of diseases using datasets and deep learning models among various hospitals, aiming to improve the accuracy of collaborative diagnosis, protect patients' privacy, and to save transmission cost.

Federated learning is first introduced by Google in 2016 to aggregate local training on mobile devices. Yang et al. [19] proposed secure federated learning to provide a solution for data isolated islands and strengthen data privacy and security. Federated learning is widely learned in lots of fields including the medical domain [20], [21]. Sheller et al. [22] studied federated learning in medicine and proposed a method to facilitate multi-institutional collaborations without sharing patient data. It proved that federated learning outperforms other collaborative learning methods such as institutional incremental learning when the federated learning datasets are independent identically distributed. Sattler et al. [23] further studied the robust federated learning from non-IID datasets. There are also federated learning based AI systems designed for COVID-19 detection using CT imaging. Kumar et al. [24] proposed blockchain-based federated learning with a data normalization technique that deals with the heterogeneity of data gathered from different hospitals and Capsule Network-based segmentation and classification. Xu et al. [25] implemented a federated learning based Unified CT-COVID AI Diagnostic Initiative which improved model generalization without data sharing.

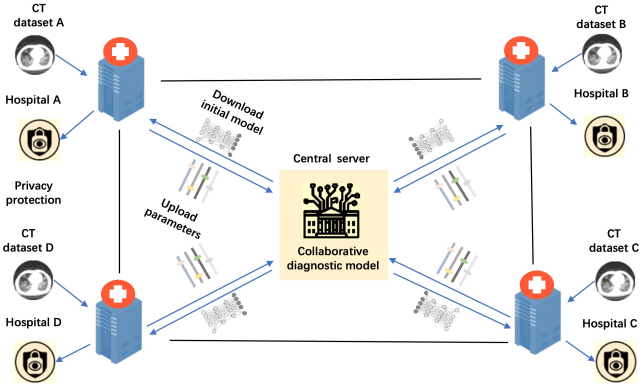


Fig. 1: The architecture of multicenter federated Learning.

Indeed, a typical federated learning framework is based on data parallelism, and the adopted parallel gradient descent strategy is to compute the local gradient on the dataset of each client. Subsequently, the weighted sum of the gradients of each client is collected as the gradient for updating collaborative model. The collaborative model is preserved by a trusted third-party institution, central server, in which the model is broadcasted to all clients at each round of iteration. After the gradients of this round of iteration are computed on each client, the central server aggregates gradients information to update the parameters of collaborative model.

In the multicenter federated learning framework as shown in **Fig.** 1, hospitals could download the code and train a local model on the basis of the initial model, and transfer the encrypted model parameters back to the central server with federated model. The central server combines the contributions shared from all of the hospitals and conduct collaborative disease diagnosis.

To further improve labeling and training efficiency, we propose to integrate active learning into multicenter federated learning. In fact, there are few studies on the topic about the integration of federated learning and active learning. Goetz et al. [26] proposed an active federated learning approach to minimize the model transmission costs by selecting clients with a probability conditioned on the current local model and local data.

B. Active Learning

Building large datasets to train a satisfactory model in a certain field is a time-consuming and labor-intensive task. For this reason, the research of active learning keeps attracting interests and attention. In general, a typical active learning framework consists of a method to evaluate the informativeness of each un-annotated data point, tied heavily to the choice of query type, after which all chosen data-points are required to be annotated. Once new annotations have been acquired, the active learning framework must use the new data to improve the model. This is normally done by either retraining the entire model using all available annotated data, or by fine-tuning the network using the most recently annotated data-points. Using this approach, state-of-the-art performance can be achieved using fewer annotations for several disease diagnosis tasks, thus widening the annotation bottleneck and reducing the costs associated with developing deep learning enabled systems from un-annotated data.

Before the emerge of deep neural networks, a host of active learning methods [27]–[31] were proposed and proved to be effective in cutting the labeling cost. However, not all of these methods can be directly used in deep learning. Yang et al. [32] proposed a suggestive deep active learning framework for CT scan image segmentation. The framework is based on a fully convolutional network architecture and selects the informative samples based on uncertainty. Zhou et al. [33] combined deep active learning with transfer learning and proposed an active fine-tuning framework called AIFT. However, this framework is only suitable for binary classification tasks. They further improved the AIFT and proposed a superior approach AFT* in [34] for adapting to multi-class cases. AFT* is also used in a carotid intimamedia thickness video interpretation task [35]. Yoo et al. [36] proposed a method to learn the loss of classification. They attached a loss prediction module to the target classifier model and jointly trained them in each round of active learning. OMedAL [37] was one of the state-of-the-art active learning frameworks for medical image analysis, which selected the most distant data points from the data distribution centroid in embedding space. The above-mentioned approaches only considered a single sampling strategy in the active learning process. The curriculum learning method [38] was proposed to select samples by combining domain-specific prior knowledge and self-paced uncertainty. Yuan et al. [39] proposed a multiple criteria deep active learning framework. However, due to excessive computational consumption, they could only utilize this method on a relatively shallow neural network.

Recently, adversarial learning was also integrated into deep active learning to learn adaptive and effective sampler. Sinha et al. [40] trained a variational auto-encoder to learn a latent space, and learned a discriminator to sample data samples with higher uncertainty.

III. METHODS

The heterogeneity and isolation of medical data of different medical institutions hinder the training of robust and generalized AI models, which is of great importance for multicenter clinical practices. In addition, under the assumption of independent identical distribution, it is reasonable to integrate the gradient information of all hospitals to update the collaborative diagnostic model according to the size of the medical dataset of each hospital. However, for multicenter federated learning, the medical dataset of each hospital are usually non-independently identically distributed among themselves. Thus it is difficult to guarantee that the final collaborative diagnostic model satisfies the requirements of all hospitals.

Inspired by federated learning and pool-based active learning methods, we propose two decentralized federated active learning methods (LEFAL and TEFAL) that introduce active learning to expand the dataset to make the each medical dataset more informative, which alleviates the negative impact of non-independently identically distribution and improves labeling and training efficiency in multicenter collaborative disease diagnosis with privacy protection.

The dscription of all symbols in the manuscript as shown in **Table** I.

A. Labeling-Efficient Federated Active Learning

In LEFAL, the local active learning process T is as follows:

- The Θ_r^k in federated round r is trained on D_r^k.
- For hospital C_k in federated round r, a hybrid sampling strategy is proposed to select an informative data subset $X_r^{k,s}$ from the unlabeled local data pool U_r^k and the oracle is requested to provide ground-truth annotations $Y_r^{k,s}$.
- The data-annotation pairs $(X_r^{k,s}, Y_r^{k,s})$ are removed from the unlabeled local data pool, and they are added to the local annotated dataset D_r^k to generate dataset D_{r+1}^k.

It is well known that the informativeness of the dataset for the model can be characterized by the uncertainty and diversity of the sample.

For data uncertainty, the data samples with high uncertainty always have a high probability to locate near the decision boundary of a learning model [41]. Therefore, we design a loss-prediction module to characterize the uncertainty of data samples by the loss value of unlabeled samples as one of the criteria for sample selection with hybrid sampling strategy. The loss-prediction module Θ_{loss}^r is jointly trained with the Θ_r^k to predict the local model loss of samples in D_r^k, which indicates

TABLE I: THE DESCRIPTION OF SYMBOLS

Symbol	Description
T	Local active learning process
r	Number of federated learning rounds
D_r^k	Labeled dataset of the k-th client in the federated round r
$X_r^{k,s}$	The data subset selected from the k-th client in federated round r
$Y_r^{k,s}$	Ground-truth annotations of $X_r^{k,s}$
U_r^k	Unlabeled dataset of the k-th client in the federated round r
Θ_r^k	The local model of the k-th client in federated round r
Θ_{loss}^r	Loss-prediction module in the federated round r
Θ_{disc}	A small network as a discriminator
L	The loss function of the loss-prediction module
d	The euclidean distance function
$\rho(x)$	The rank function
l^p	The loss pair (l_i, l_j)
$\hat{l}^p$	The predicted loss pair $(\hat{l}_i, \hat{l}_j)$
x_u^k	The samples to be selected
x_i^k	The samples in local dataset
ζ	A positive constant which represents the margin between larger loss and smaller loss
α,β	Constant between 0 and 1
b	The local labeling budget
f	The feature embedding
$\mathcal{I}^k$	The quantification of the diversity of the samples
$\mathcal{L}^k$	The quantification of the uncertainty of the dataset
$\mathcal{E}^k$	The average prediction loss
F^k	An evaluation vector
C_k	The k-th client or hospital
$\varkappa^k$	The k-th client or hospital value score

the data uncertainty. The loss function of the loss-prediction module is designed as follows:

$$L\left(\hat{l}^p, l^p\right) = \max(0, -Z(l_i, l_j) \cdot (\hat{l}_i - \hat{l}_j) + \zeta)$$
$$s.t. \quad Z(l_i, l_j) = \begin{cases} 1, l_i > l_j \\ -1, otherwise \end{cases} \tag{1}$$

Among them, l^p denotes the loss pair (l_i, l_j) and $\hat{l}^p$ denotes the predicted loss pair $(\hat{l}_i, \hat{l}_j)$, and ζ denotes a positive constant which represents the margin between larger loss and smaller loss. When $l_i > l_j$, the expected loss-prediction module should give the prediction $\hat{l}_i > \hat{l}_j$ and the difference between them exceeds ζ, otherwise a penalty is imposed to Θ_{loss}^r to update its weights. Thus we can use the output of Θ_{loss}^r as a quantification to evaluate the uncertainty of $x_u^k \in U_r^k$:

$$uncertainty(x_u^k) = \Theta_{loss}^r(x_u^k) \tag{2}$$

For data diversity, data samples with the same type of information generally show a certain clustering pattern in the feature space [42]. Accordingly the newly selected data samples for active learning should be kept at a considerable distance from the existing data samples to ensure new data can indeed introduce effective diversity information. Therefore, we propose to use the average Euclidean distance between the unlabeled sample and the samples in local dataset as a quantification to evaluate the gain in sample diversity:

$$diversity(x_u^k) = \frac{1}{||D_{r-1}^k||} \sum_{i=1}^{||D_{r-1}^k||} d\left(f(x_i^k), f(x_u^k)\right) \tag{3}$$

Among them, $x_u^k \in U_r^k$ denotes the samples to be selected, $x_i^k \in D_{r-1}^k$ denotes the samples in local dataset, f denotes the feature embedding, and d denotes the Euclidean distance function. In addition, considering the computational efficiency, we use the mean value of the local dataset in the feature space as the estimate:

$$diversity(x_u^k) = d\left(\left(\frac{1}{||D_{r-1}^k||} \sum_{i=1}^{||D_{r-1}^k||} f(x_i^k)\right), f(x_u^k)\right) \tag{4}$$

To effectively integrate data uncertainty with data diversity, a discretization method is utilized to convert the evaluated values into scale-free ranks. The data samples are sorted in the ascending order according to predicted model loss and computed embedding distance. Hence, the rank function $\rho(x)$ means to find the position of a certain data sample in the ascending sorted list of the evaluated metrics. We can have $\rho_u(x)$ for data uncertainty and $\rho_d(x)$ for data diversity respectively. Since the ranks are scale-free, a hybrid rank $\rho_h(x)$ can be achieved by weighted summing up $\rho_u(x)$ and $\rho_d(x)$. In addition, we use α and β to adjust the relative influence of the two metrics in the hybrid sampling:

$$\rho_h(x) = \alpha\rho_u(x) + \beta\rho_d(x) \tag{5}$$

Thus, the oracle is requested to label x^* which maximizes the value of $\rho_h(x)$ in each sampling iteration. The sampling iterations will repeat until local labeling budget b is exhausted. The optimized data subsets construct the most representative local distributions for global aggregation will be:

$$x^* = \arg\max_{x \in U_r^k} \rho_h(x) \tag{6}$$

B. Training-Efficient Federated Active Learning

TEFAL is designed for selecting most informative hospitals in each federated round r to boost the training process and reduce model transmission costs. The optimized hospital subset is selected based on the value evaluation of the hospitals.

We consider that the higher the training performance in a round of federated learning iterations, the more local models that can contribute to the convergence of the global model. It is obvious that the value of a hospital is directly proportional to its contribution to the global model under current round of aggregation. Moreover, it is believed that the quality of the dataset directly affects the training effect and performance of a learning model. According to previous analysis, the quality of the dataset of a hospital can be evaluated by the size of

datasets, the data diversity of samples, and the percentage of hard samples near the decision boundary of the model.

For hospital C_k, there is a local dataset D^k. According to the previous analysis in LEFAL, we also adopt the Euclidean distance $\mathcal{I}^k$ as a quantification of the diversity of the samples and the training loss $\mathcal{L}^k$ as a quantification of the uncertainty of the dataset. In addition, we divide a small portion of the local dataset into an exploration dataset and use the average prediction loss $\mathcal{E}^k$ of the model on exploration set as a quantification indicator of the optimization progress of the local model as a way to assess the contribution of the local model to the global model. In summary, we construct an evaluation vector F^k for each hospital according to the above four metrics:

$$F^k = (||D^k||, \mathcal{I}^k, \mathcal{L}^k, \mathcal{E}^k) \tag{7}$$

The central server collects quadruples F^k from all the hospitals for evaluation. A small network Θ_{disc} is trained in the server as a discriminator to make decisions whether a hospital should be selected for current federated learning round. Inspired by adversarial active learning, we keep the discriminator part and cut the generator off. The attribute values in evaluation quadruples are normalized to $[0, 1]$ and they are input to the discriminator. The task for Θ_{disc} is to identify the value of hospitals. The hospital value score $\varkappa^k$ is evaluated with the output of the discriminator, thus the values can be used to construct a selection metric:

$$\varkappa^k = \Theta_{disc}\left(normalize(F^k)\right) \tag{8}$$

For the obtained selection metric $\{\varkappa_k\}_{k=1}^K$, we can either use it directly as the aggregation weight for the original federation learning, or we can take a top-k approach to select appropriate hospitals for aggregation to reduce the transmission costs.

IV. ARCHITECTURE

A. Labeling-Efficient Federated Active Learning

In the proposed LEFAL method, active learning is integrated into a standard federated learning framework. The LEFAL framework is illustrated in **Fig.** 2. and the detailed working process is described in **Alg.** 1. We will discuss the workflow of the LEFAL in the aspects of the central server and the hospitals.

In terms of the server, it collects local models from the hospitals in each federated round and aggregates these local models to achieve the global model. Subsequently the server sends the parameters of collaborative diagnostic model to each hospital.

For each hospital in a federated round, an informative data subset is selected with the hybrid sampling strategy described in **Alg.** 2. The LEFAL then queries the ground-truth annotations (segmentation masks or classification labels) from the oracle (radiology experts in different hospitals). The data-annotation pairs are removed from the unlabeled local data pool, and added to the local annotated dataset. After that, the loss-prediction module are jointly trained on the local labeled dataset. The detailed information of the loss-prediction module can be found in **Fig.** 3.

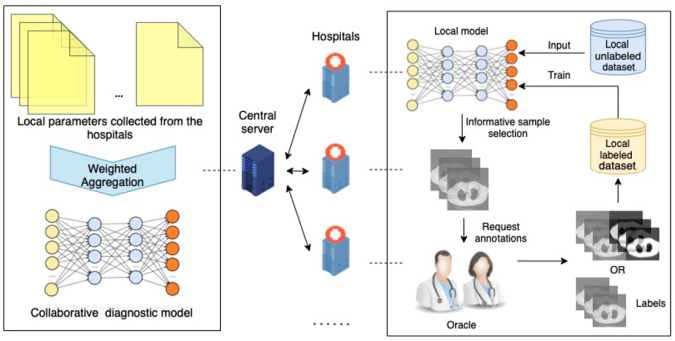


Fig. 2: The architecture of Labeling-Efficient Federated Active Learning.

Algorithm 1 Labeling-Efficient Federated Active Learning

Input:
Participated hospital group $C = \{C_k\}_{k=1}^K$
Unlabeled local data pool $\{U_r^k\}_{k=1}^K$
Local models $\{\Theta_r^k\}_{k=1}^K$
Local loss-prediction module Θ_{loss}^r
Labeling budget b in each federated round r
Global model Θ_{global}
Output:
Well-trained global model Θ_{global}^*
Local labeled datasets $\{D_r^k\}_{k=1}^K$
for each C_k in C **do**
  Randomly label a small subset of X^k to initialize D_0^k
end for
repeat
  for each C_k in C **do**
    Jointly train Θ_r^k and Θ_{loss}^r on D_r^k
    Sample an informative data subset $X_r^{k,s}$ of labeling budget b from U_r^k according to **Alg.** 2 and query labels from the oracle
    $D_{r+1}^k \leftarrow D_r^k \cup \left(X_r^{k,s}, Y_r^{k,s}\right)$, $U_{r+1}^k \leftarrow U_r^k \backslash X_r^{k,s}$
    Upload local model parameters ω_r^k to the server
  end for
  Aggregate the local model parameters $\{\omega_r^k\}_{k=1}^K$ and update the global model Θ_{global}
until *global task performance is satisfied*
return Θ_{global}^*, $\{D_r^k\}_{k=1}^K$

In **Fig.** 3, the process of the hybrid sampling strategy is described and the classification task is presented as an example. "Block" refers to the feature extraction module composed of some Convolutional layers and BN layers in the neural network model. For example, in the classification experiment, the "Block" is the residual block in the Resnet18 [43] network model. These residual blocks are stacked by convolutional layers and BN layers. "Block 1", "Block 2" refers to the different residual blocks. The loss prediction module outputs the predicted scalar loss values by fusing features from different layers. In the model learning process, the output predictions are compared with ground-truth annotations to compute training loss. The training loss is taken as ground-truth labels for the loss-prediction module, which is attached to obtain feature maps as input. The feature maps are processed with global pooling layers and fully-connected layers, then they are concatenated into a feature vector for loss prediction. At the same time, the feature embeddings of input data are extracted by the appointed embedding extraction layer of the

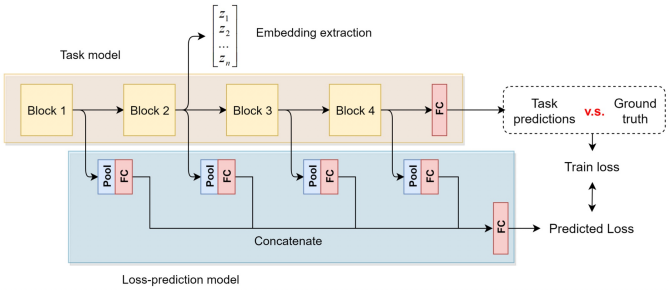


Fig. 3: The model to support the hybrid sampling strategy.

model. The attached loss-prediction module can be jointly trained with the model and the embedding extraction process is simultaneously accomplished. That is to say, there will not be extra training cost.

Algorithm 2 Hybrid Sampling Strategy

Input:
Unlabeled data pool U
Labeled dataset D
Model Θ
Loss-prediction module Θ_{loss}
Labeling budget b
Output:
Informative data subset X^s

repeat
  Update the embedding centroid of labeled dataset D
  for each x_u in U **do**
    Predict model loss by Θ_{loss}
    Calculate the *uncertainty* of sample $uncertainty(x_u)$
    Extract the embedding of x_u by feature extraction layers of Θ_{task}
    Calculate the *diversity* of sample $diversity(x_u)$
  end for
  for each x_u in U **do**
    Transfer the *uncertainty* and *diversity* to scale-free rank value
    $\rho_h(x_u) \leftarrow \alpha\rho_u(x_u) + \beta\rho_d(x_u)$
  end for
  Sample the most informative data sample x^* according to ρ_h
  Add x^* to X^s
until *sampling budget b is exhausted*
return X^s

The proposed hybrid sampling strategy has two main advantages. It simultaneously considers data uncertainty and diversity, which could measure the data informativeness comprehensively. In addition, the proposed hybrid sampling strategy is task-agnostic, which is compatible with both segmentation and classification tasks. The loss-prediction module and embedding extraction layer are adaptive in various learning tasks.

For data uncertainty, we propose to measure it with a loss-prediction module. Intuitively, the model will have higher loss when it meets hard samples with more uncertainty. In other words, the data uncertainty can be achieved if the loss can be predicted before actual training. To meet this goal, we tailored design a loss-prediction module shown in **Fig.** 3 to predict the task model loss. The task loss prediction can be a special regression problem, where normal mean square error loss is not feasible because of the rapid fluctuation of the task loss. To address this problem, the loss function for the loss-prediction module is specially designed in **Eq.** 7. While in the sampling stage, the loss-prediction module will predict the task model loss for each unlabeled data samples. The data samples with higher predicted loss are considered more informative.

For data diversity, we propose to measure it by computing the distances of feature embeddings in the latent space. Traditional method computes the average distance from one unlabeled data sample to each labeled data sample in the latent space, which results in low computational efficiency. To simplify the calculation, we assume the latent space of the embedding function is always Euclidean. Thus, the calculation is equivalent to computing the distance from one unlabeled data sample to the centroid of all the labeled data samples in the embedding space. This assumption effectively cut down the time complexity for embedding distance computation from $O(n^2)$ to $O(n)$. Each time the most informative data sample x^* is selected by the proposed hybrid sampling strategy, the centroid of the labeled samples is also updated to include the newly selected data sample.

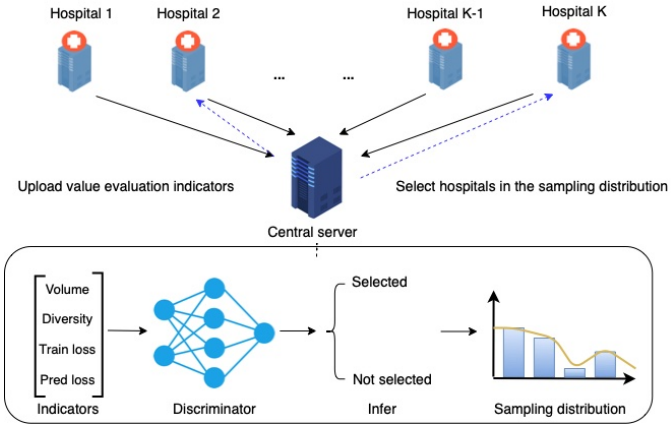


Fig. 4: The architecture of Training-Efficient Federated Active Learning.

B. Training-Efficient Federated Active Learning

Federated learning allows the collaborative diagnostic model to be trained without centralizing hospital data by transmitting global model to hospitals, calculating gradients locally, then averaging these gradients. Usually larger models are implemented with higher accuracy performance. The model transmissions between hospitals and the server occupy the hospitals' bandwidth. It is of great importance to minimize these transmission costs in federated learning. Different from the informative data sample selection in LEFAL, TEFAL heuristically select an optimal hospital subset in each federated round for not only lower transmission cost but also faster global model convergence.

In the federated setting, only a summary of the hospitals is available to central server for the sake of privacy protection. As illustrated in **Fig.** 4, the hospitals of TEFAL return value evaluation indicators to the server after training process. Besides standard weight aggregation steps, the central server of TEFAL collects quadruples from the hospitals including the information of training loss, data volume, predicted loss, and data diversity. A discriminator is trained to decide the

Algorithm 3 Training-Efficient Federated Active Learning

Input:
Participated hospital group $C = \{C_k\}_{k=1}^{K}$
s $\{\Theta_r^k\}_{k=1}^{K}$
Local loss-prediction modules $\{\Theta_{loss}\}$
Local labeled datasets D^k
Global model Θ_{global}
Discriminator Θ_{disc}
Hospital selection budget b in each federated round
Output:
Well-trained global model Θ^*_{global}

repeat
  for each C_k in C **do**
    Jointly train Θ_r^k and Θ_{loss} on D_r^k
    Compute value evaluation quadruple $\digamma^k = \left(||D^k||, \mathcal{I}^k, \mathcal{L}^k, \mathcal{E}^k\right)_r$
    Normalize the quadruple to $[0, 1]$
    Infer value evaluation $\varkappa_r^k$ with Θ_{disc}
  end for
  Select b valuable hospitals in C to form a hospital subset S_r
  Aggregate the Θ_r^k according to S_r and update the global model Θ_{global}
  Train Θ_{disc} with the selection result
until *global task performance is satisfied*
return Θ^*_{global}

probability of whether a hospital should be selected in the next federated round. The input of the discriminator is 4-dimensional vectors, and we use a small neural network with two ReLU activated fully connected layers and a Sigmoid activated output. The output of this discriminator constructs a sampling distribution for hospital selection. The hospitals participating in the next federated round are selected in this sampling distribution. The server aggregates the local parameters of the selected informative hospitals in one federated round. The **Alg.** 3 describes the detailed learning process of TEFAL.

V. EXPERIMENTAL EVALUATION

A. Datasets setup

For the qualitative and quantitative analysis of proposed methods, we conduct experiments on lung CT dataset and Hyper-Kvasir. The lung CT dataset was constructed from cohorts from China Consortium of Chest CT Image Investigation for COVID-19(for simplicity, we call it CC-CCII [13]). Hyper-kvasir [14] is a gastrointestinal endoscopy dataset, collected during real gastroscopy and colonoscopy at Bærum hospital in Norway and partly labeled by experienced gastrointestinal endoscopists. CC-CCII contains 617,775 axial slices of CT scans from 4,154 human subjects. In the first version of the dataset, only patient-level labels were provided for weak supervised learning. The latest version furnishes lesion slice information and 750 segmentation masks to meet the demands for further researches. Hyper-Kvasir can be split into four distinct partsLabeled image data, unlabeled image data, segmented image data, and annotated video data. In this paper, we only use labeled image data and segmented images. The labeled image data includes 10662 image data, including 23 types of gastrointestinal diseases. The segmented images are 1000 images and segmentation masks from polyps.

In the experiment, we use the dataset for the classification task and segmentation task, respectively. Image segmentation and image classification are two separate tasks for disease diagnosis. The segmentation task is a semantic segmentation of the lesion area, in which the medical image is used as input and a class is assigned to each pixel in the image. The classification task is to determine whether a patient has a disease or not or what kind of disease, in which a patient's medical image is used as input to make a classification according to the entire image. For CC-CCII, in order to complete the classification task, we divided the samples in the CT slice dataset into 3 categories, including *Novel Coronavirus Pneumonia* due to SARS-CoV-2 virus infection, *Common Pneumonia*, and *Normal Control*. The human subjects were considered clinically appropriate for chest CT scans while the association of age and gender were not taken into consideration. Since the majority of the CT slices are classified into NC, we manually control the volume of NC slices to 50,000. For the segmentation task, we segmented 750 CT slices of COVID-19 patients into *Background*, *Lung Fields*, *Ground Glass Opacity* (GGO) and *Consolidation* (CL). For Hyper-Kvasir, due to the extreme class imbalance in the labeled image data, we selected Polyphs, Esophagitis and Z-line from 23 gastrointestinal diseases for classification experiments.In the segmentation task, we segmented 1000 polyphs images into background and foreground.The details of the CC-CCII dataset and Hyper-Kvasir dataset are shown in **Table** II.

TABLE II: STATISTICS OF DATASET

Task	Label	Volume
CC-CCII Segmentation	Background/Lung/GGO/CL	750
CC-CCII Classification	Novel Coronavirus Pneumonia	21872
	Common Pneumonia	36894
	Normal Control	50000
Hyper-Kvasir Segmentation	Background/Foreground	1000
Hyper-Kvasir Classification	Polyphs	1028
	Esophagitis	757
	Z-line	932

B. Architectures and Optimization Hyperparameters

We carry out medical image segmentation and classification experiments on the CC-CCII and Hyper-Kvasir respectively to verify the effectiveness of the proposed methods. For the segmentation task, we adopt U-Net [44] as and global model for federated learning. Whereas for the classification task, ResNet18 [43] model is used. We uniformly resize the medical images to 224 $\times$ 224 as the input. The number of clients participating in the multicenter collaborative disease diagnosis is set to 5. The batch size is set to 12. The number of local epochs is set to 10. We use the Adam optimizer with a learning rate 1e-4 for both segmentation and classification tasks. The Adam weight decay is set to 5e-4 and the Adam momentum is set to 0.9. In addition, the loss prediction module uses the SGD optimizer with a learning rate 1e-3. The SGD weight decay is set to 5e-4 and the Adam momentum is set to 0.9.

C. Evaluation Metrics

To quantitatively compare the performance of different models, we use the corresponding evaluation metrics for each task. For the segmentation task, we adopt mean intersection over union (Mean-IOU), while for the classification task, we use accuracy, precision, recall and F1-score:

$$\begin{aligned} IOU &= \frac{\hat{y} \cap y}{\hat{y} \cup y} \\ DICE &= \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|} \\ Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Recall &= \frac{TP}{TP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ F1 &= \frac{2 \times Precision \times Recall}{Precision + Recall} \end{aligned} \tag{9}$$

D. Experimental Results

1) Federated Diagnosis: To confirm that collaborative diagnosis based on federated learning has improved performance over learning a diagnostic model on a single hospital, we design a set of basic experiments on CC-CCII and Hyper-Kvasir. We manually allocate different portion of training data to each hospital and set aside 10% of the data as the test set, so as to compare the effect of different sizes of datasets on model performance. In addition, we aggregate the local model weights of each clients to form a simplified federated learning model as an illustration of the effect of federated learning for the classification and segmentation tasks. And we iterate the model 50 times for single local model on the segmentation task and 20 times for the classification task, respectively. Similarly, we train 50 federated rounds for the federated learning framework on the segmentation task, while 20 federated rounds were trained on the classification task.

As demonstrated in **Table** III and **Table** IV, the performance of collaborative diagnosis model with federated learning is better than the local model of each hospital with limited data. Moreover, the collaborative diagnosis model with federated learning is slightly better than the virtual hospital using 90% of the total amount of accumulated data, which is impossible in reality due to the privacy protection of patients from each hospital. The results in **Table** III and **Table** IV are presented to prove the effectiveness of proposed federated learning method without the transmission of raw data.

2) Labeling-Efficient Federated Active Learning: To verify the effectiveness of the hybrid sampling strategy proposed in LEFAL, we compare the performance of the model under the same conditions with different sampling methods. And we use the 5-fold cross-validation method to ensure the stability of the experiments. Moreover, to explore the effectiveness of the quantitative metrics characterizing sample diversity and uncertainty in the hybrid sampling strategy, we conducted a set of simultaneous ablation experiments:

- Random sampling strategy (RAND): randomly selecting samples to be manually labeled in each federated round.

TABLE III: THE COMPARISON BETWEEN FEDERATED LEARNING AND SINGLE INSTITUTION LEARNING ON SEGMENTATION TASK

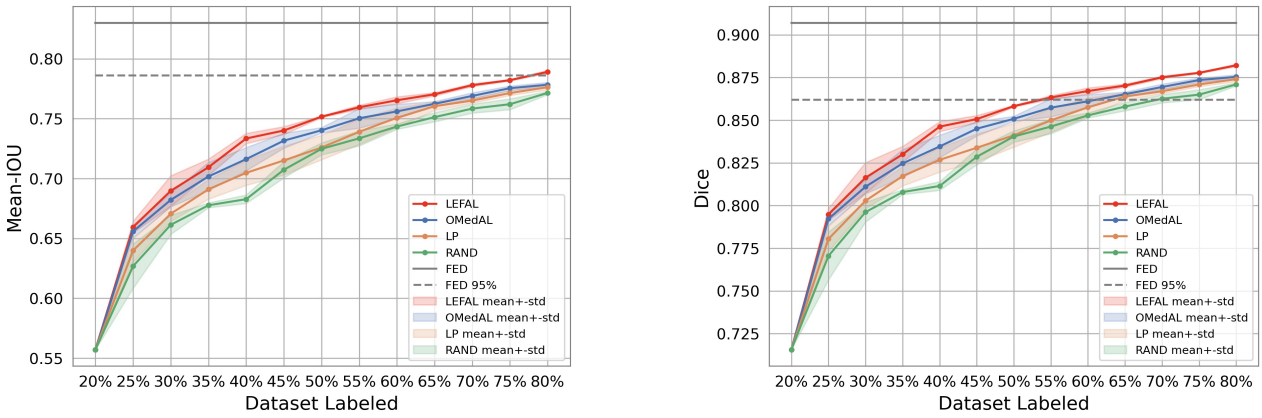
Datasets	Institution	Data Portion	IOU	Dice
CC-CCII	Hospital1	5%	0.755	0.860
	Hospital2	10%	0.758	0.862
	Hospital3	15%	0.765	0.867
	Hospital4	20%	0.772	0.871
	Hospital5	40%	0.793	0.885
	Virtual Hospital	90%	0.827	0.905
	FL	90%	**0.830**	**0.907**
Hyper-Kvasir	Hospital1	5%	0.717	0.835
	Hospital2	10%	0.721	0.838
	Hospital3	15%	0.735	0.847
	Hospital4	20%	0.752	0.858
	Hospital5	40%	0.773	0.872
	Virtual Hospital	90%	0.804	0.891
	FL	90%	**0.806**	**0.893**

TABLE IV: THE COMPARISON BETWEEN FEDERATED LEARNING AND SINGLE INSTITUTION LEARNING ON CLASSIFICATION TASK

Datasets	Institution	Data Portion	Acc	Pre	Recall	F1
CC-CCII	Hospital1	5%	0.941	0.940	0.939	0.939
	Hospital2	10%	0.946	0.951	0.944	0.947
	Hospital3	15%	0.948	0.950	0.953	0.951
	Hospital4	20%	0.955	0.952	0.956	0.954
	Hospital5	40%	0.973	0.972	0.971	0.972
	Virtual Hospital	90%	**0.977**	**0.973**	0.978	0.975
	FL	90%	0.976	0.972	**0.978**	**0.975**
Hyper-Kvasir	Hospital1	5%	0.754	0.753	0.749	0.751
	Hospital2	10%	0.778	0.779	0.776	0.776
	Hospital3	15%	0.782	0.784	0.781	0.782
	Hospital4	20%	0.797	0.800	0.796	0.799
	Hospital5	40%	0.819	0.821	0.818	0.820
	Virtual Hospital	90%	0.839	0.835	0.840	0.833
	FL	90%	**0.840**	**0.833**	**0.841**	**0.836**

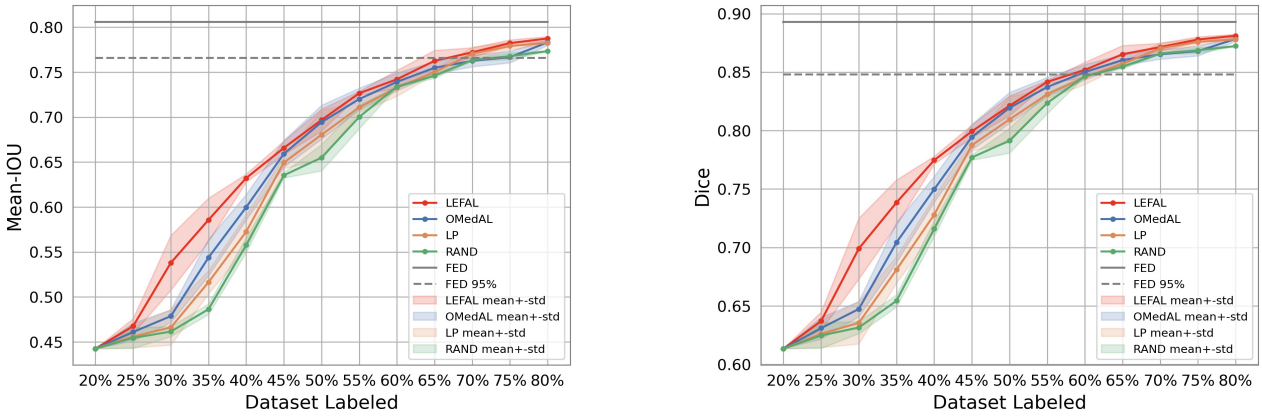
- Loss-prediction based sampling strategy (LP): selecting samples with larger uncertainty quantification based only on the loss-prediction module.
- OMedAL [37]: selecting samples with larger diversity quantification based only on the Euclidean distance.
- Original federated learning (FED): without using active learning method.

The comparison results of LEFAL and the state-of-the-art active learning methods on the segmentation task and the classification task of COVID-19 diagnosis and gastrointestinal disease are shown in **Fig.** 5 to **Fig.** 8, respectively. Initially, 20% of the training set is labeled randomly. And then, in each federated round, an additional 5% of the dataset will be labeled. In each federated round, the clients conduct 1 active cycle. The gray solid lines show the performance of FED running for 50 rounds. Since training on complete datasets obviously outperforms training on informative subsets, we plot gray dashed lines which illustrate 95% performance on the classification task and segmentation task. As shown in **Fig.** 5 to **Fig.** 8, the performance of the model is greatly improved compared to the original federated learning framework after expanding the dataset by active learning sampling.



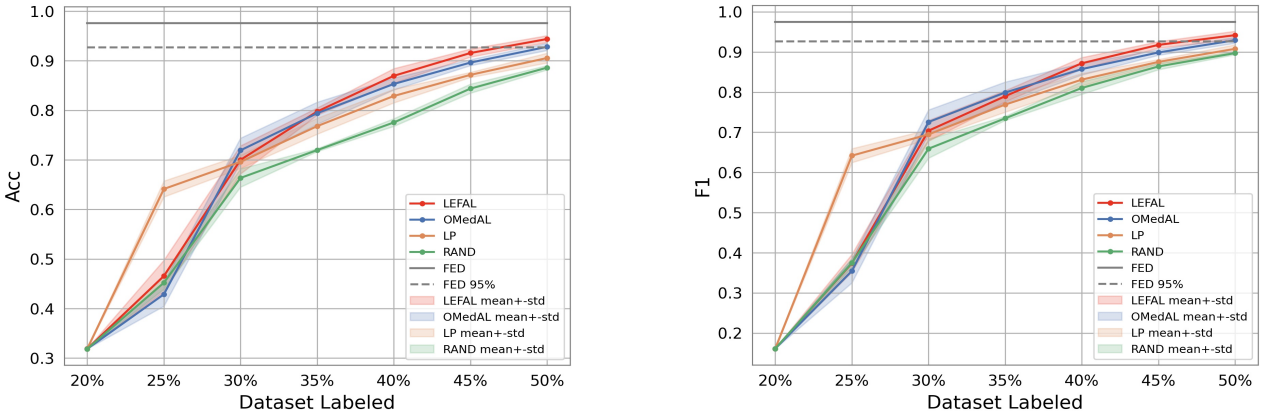
(a) (b)

Fig. 5: The Comparison between LEFAL and other methods in segmentation task on CC-CCII dataset.(a) The Mean-IOU of LEFAL and other methods.(b) The Dice of LEFAL and other methods.



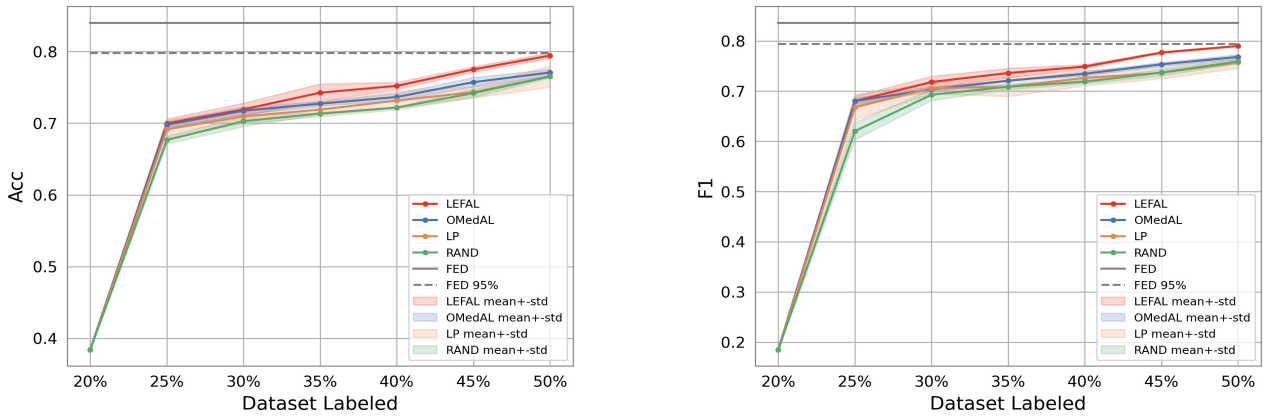
(a) (b)

Fig. 6: The Comparison between LEFAL and other methods in segmentation task on Hyper-Kvasir dataset.(a) The Mean-IOU of LEFAL and other methods.(b) The Dice of LEFAL and other methods.



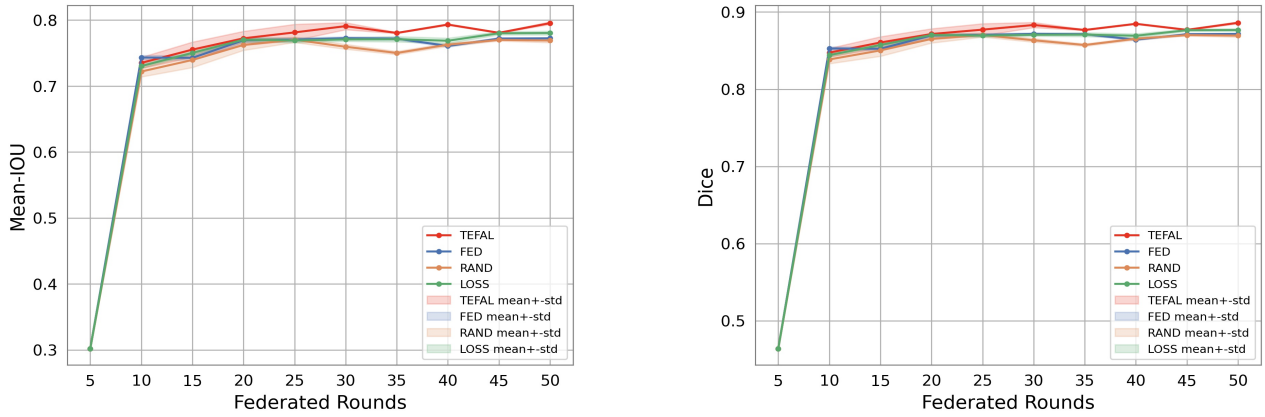
(a) (b)

Fig. 7: The Comparison between LEFAL and other methods in classification task on CC-CCII dataset.(a) The Accuracy of LEFAL and other methods.(b) The F1 value of LEFAL and other methods.



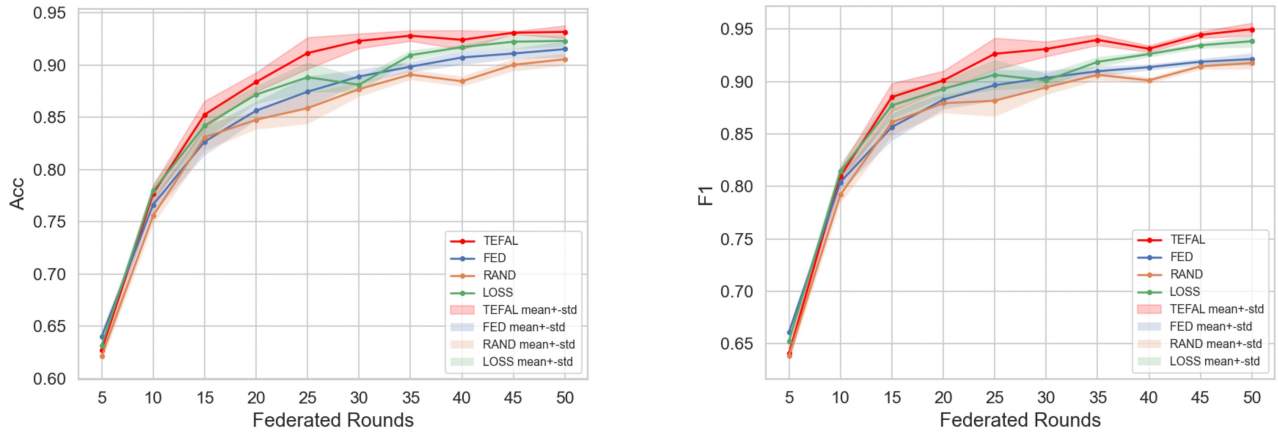
(a) (b)

Fig. 8: The Comparison between LEFAL and other methods in classification task on Hyper-Kvasir dataset.(a) The Accuracy of LEFAL and other methods.(b) The F1 value of LEFAL and other methods.



(a) (b)

Fig. 9: The Comparison between TEFAL and other methods in segmentation task on CC-CCII dataset.(a) The Mean-IOU of TEFAL and other methods.(b) The Dice of TEFAL and other methods.



(a) (b)

Fig. 10: The Comparison between TEFAL and other methods in classification task on CC-CCII dataset.(a) The Accuracy of TEFAL and other methods.(b) The F1 value of TEFAL and other methods.

The proposed hybrid selection strategy will pick data samples with both high uncertainty and diversity in each round of active learning. However, the LP method only considers the uncertainty of the samples according to prediction loss. Thus it achieves good performance at the early stage of training since it picks hard samples near the decision boundary to accelerate the convergence. As the training goes on, our proposed method outperforms the LP method due to the hybrid sampling strategy. In addition, the LEFAL model proposed in this paper outperforms other methods in both segmentation and classification tasks.

3) Training-Efficient Federated Active Learning: Moreover, to verify the efficiency of the TEFAL proposed in this paper, we also compared it with other approaches for selecting clients:

- Random sampling strategy (RAND), which random selects n clients in each federated round.
- Loss-based sampling strategy (LOSS), which selects n clients with highest local training loss in each federated round.
- Original federated learning (FED), which aggregates the training results from all clients in each federated round.

The comparison results of TEFAL and other approaches on the segmentation task and the classification task are shown in **Fig.** 9 and **Fig.** 10, respectively. It demonstrate the experimental comparison between TEFAL and the state-of-the-art client selection methods in federated learning on segmentation task and classification for COVID-19 diagnosis respectively.

The experimental results reveal that randomly selecting client subset for lower transmission cost is harmful to the diagnostic performance in both tasks. The loss-based selection is beneficial and TEFAL even outperforms this method. The TEFAL gains higher convergence speed with informative client subset selection.

E. Discussion

It is noteworthy that for the segmentation task, the LEFAL almost achieves 95% performance using 65% labels of the whole dataset on the Hyper-Kvasir dataset. For the classification task, the LEFAL achieves 95% accuracy and F1-score with 50% labels on both datasets. Therefore, using a focused sampling strategy to select valuable samples is effective in improving the quality of the dataset compared to uniform random sampling, thus improving the model performance. Moreover, after quantifying the value of samples based on metrics, it can also greatly reduce the cost of labeling invalid samples and improve the efficiency of active learning.

Federated Learning methods select client to reduce communication cost. Thus high performance are selected in the proposed TEFAL method to facilitate federated training and reducing communication. The performance of local model on each client is directly related to the quality of its training dataset, and the quality of a dataset can be described through data diversity and model uncertainty. Moreover, local training loss and predictive loss on each client are also critical factors. Therefore, data diversity, model uncertainty, local training loss and prediction loss are used as features to calculate the weight of clients to ensure that high-performance clients are selected in each communication round. As a result, the proposed TEFAL method outperforms the state-of-the-art methods.

To verify the uncertainty of our models, we introduce Monte-carlo dropout into model reasoning. The verification results are shown in **Table** V and **Table** VI.

TABLE V: MONTE-CARLO DROPOUT ON SEGMENTATION TASK

Sample ID	Order	Segmentation(IOU)	Variance
Polyps/24849	1	0.7813	1.2491e-5
	2	0.7789	
	3	0.7721	
	4	0.7748	
	5	0.7806	

TABLE VI: MONTE-CARLO DROPOUT ON CLASSIFICATION TASK

Sample ID	Order	Classification	Entropy
Polyps/24849	1	[1.4e-3, 9.968e-1, 1.8e-3]	3.585e-2
	2	[1.6e-3, 9.971e-1, 1.3e-3]	
	3	[1.1e-3, 9.97e-1, 1.90e-3]	
	4	[2.1e-3, 9.961e-1, 1.9e-3]	
	5	[1.9e-3, 9.955e-1, 2.7e-3]	
Esophagitis/14d54	1	[0.6931, 0.0965, 0.2104]	1.127
	2	[0.7168, 0.1185, 0.1647]	
	3	[0.7586, 0.0841, 0.1573]	
	4	[0.6863, 0.1133, 0.2004]	
	5	[0.7223, 0.0855, 0.1922]	
Z-line/054eb	1	[0.3276, 0.0231, 0.6493]	1.049
	2	[0.3010, 0.0212, 0.6778]	
	3	[0.3090, 0.0199, 0.6712]	
	4	[0.2822, 0.0270, 0.6908]	
	5	[0.3776, 0.0244, 0.5980]	

VI. CONCLUSION

With the rapid development of deep learning techniques in medical diagnosis, efficient collaborative diagnosis has become an important issue to preserve data privacy and aggregate the contributions from various institutions. To address this problem, we propose the idea of federated active learning and present LEFAL method with hybrid data sampling and TEFAL method with informative client selection. The performance of these two methods are evaluated not only on the segmentation tasks but also on classification tasks for Covid-19 diagnosis and gastrointestinal disease diagnosis. The empirical results demonstrate that the proposed LEFAL method and TEFAL method outperforms the state-of-the-art methods with higher efficiency and better accuracy. Medical institutions generate a large number of unlabeled medical data everyday and the cooperation among medical institutions will be closer in the near future. Thus it is of great importance to propose practical methods for effective and efficient cooperation among medical institutions to reduce data labeling cost under the premise of protecting data privacy. Similarly, many industries, such as banks, securities and Internet companies could take advantage of the proposed LEFAL method and TEFAL method to reduce labeling cost and protect data privacy simultaneously during

multilateral cooperation. The encrypted transmission is beyond the scope of this paper, which may be the limitation of proposed methods for bank's business. However, in the joint modeling of abnormal transactions in multiple banks, the proposed method can improve the data utility when each bank only has a small amount of abnormal transaction data samples. That is to say, the proposed LEFAL and TEFAL methods do have strengths and limitations for different applications. The encrypted transmission could be one of our future works.

REFERENCES

[1] F. Shi, J. Wang, J. S. hi, and et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Reviews in Biomedical Engineering*, PP(99):1–1, 2020.

[2] X. Mei, H.C. Lee, K. Diao, and et al. Artificial intelligence for rapid identification of the coronavirus disease 2019 (covid-19). *medRxiv*, pages 2020–04, 2020.

[3] J. Lei, J. Li, X. Li, and et al. Ct imaging of the 2019 novel coronavirus (2019-ncov) pneumonia. *Radiology*, 295(1):18–18, 2020.

[4] X. Wu, M. Zhong, Y. Guo, and et al. The assessment of small bowel motility with attentive deformable neural network. *Information Sciences*, 508:22–32, 2020.

[5] M. Pei, X. Wu, Y. Guo, and et al. Small bowel motility assessment based on fully convolutional networks and long short-term memory. *Knowledge-Based Systems*, 121:163–172, 2017.

[6] S. Kashyap, H. Zhang, K. Rao, and et al. Learning-based cost functions for 3-d and 4-d multi-surface multi-object segmentation of knee mri: Data from the osteoarthritis initiative. *IEEE Transactions on Medical Imaging*, 37(5):1103–1113, 2018.

[7] D. Wang and Y. Shang. A new active labeling method for deep learning. In *Proceedings of International Joint Conference on Neural Networks*, pages 112–119, 2014.

[8] V. Nath, D. Yang, B. A. Landman, and et al. Diminishing uncertainty within the training pool: active learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2534–2547, 2021.

[9] J. Melendez, B. van Ginneken, P. Maduskar, and et al. On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis. *IEEE Transactions on Medical Imaging*, 35(4):1013–1024, 2016.

[10] F. C. Ghesu, E. Krubasik, B. Georgescu, and et al. Marginal space deep learning: efficient architecture for volumetric image parsing. *IEEE Transactions on Medical Imaging*, 35(5):1217–1228, 2016.

[11] W. Luo, A. Schwing, and R. Urtasun. Latent structured active learning. *Advances in Neural Information Processing Systems*, 26, 2013.

[12] A. Linardos, K. Kushibar, S. Walsh, and et al. ederated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports*, 12(1):1–12, 2022.

[13] K. Zhang, X. Liu, J. Shen, and et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020.

[14] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

[15] Xiaofei Wang, Lai Jiang, Liu Li, Mai Xu, Xin Deng, Lisong Dai, Xiangyang Xu, Tianyi Li, Yichen Guo, Zulin Wang, and Pier Luigi Dragotti. Joint learning of 3d lesion segmentation and classification for explainable covid-19 diagnosis. *IEEE Transactions on Medical Imaging*, 40(9):2463–2476, 2021.

[16] Cheng Xue, Lequan Yu, Pengfei Chen, Qi Dou, and Pheng-Ann Heng. Robust medical image classification from noisy labeled data with global and local representation guided co-training. *IEEE Transactions on Medical Imaging*, 41(6):1371–1382, 2022.

[17] B. Thamsen, P. Yevtushenko, L. Gundelwein, A. A. A. Setio, H. Lamecker, M. Kelm, M. Schafstedde, T. Heimann, T. Kuehne, and L. Goubergrits. Synthetic database of aortic morphometry and hemodynamics: Overcoming medical imaging data availability. *IEEE Transactions on Medical Imaging*, 40(5):1438–1449, 2021.

[18] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[19] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[20] T. Li, A. K. Sahu, A. Talwalkar, and et al. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[21] J. Xu, B. S. Glicksberg, C. Su, and et al. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

[22] M. J. Sheller, B. Edwards, G. A. Reina, and et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1):1–12, 2020.

[23] F. Sattler, S. Wiedemann, K. Müller, and et al. Robust and communication-efficient federated learning from non-iid data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2019.

[24] A. A. Kumar, R.and Khan, J. Kumar, and et al. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. *IEEE Sensors Journal*, 21(14):16301–16314, 2021.

[25] Y. Xu, L. Ma, F. Yang, and et al. A collaborative online ai engine for ct-based covid-19 diagnosis. *medRxiv*, 2020.

[26] J. Goetz. *Active learning in non-parametric and federated settings*. PhD thesis, 2020.

[27] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *Proceedings of the* 21^{st} *International Conference on Machine Learning*, page 79, 2004.

[28] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009.

[29] X. Wu, C. Chen, M. Zhong, and et al. Hal: Hybrid active learning for efficient labeling in medical domain. *Neurocomputing*, 456:563–572, 2021.

[30] X. Wu, C. Chen, M. Zhong, and et al. Covid-al: The diagnosis of covid-19 with deep active learning. *Medical Image Analysis*, 68:101913, 2021.

[31] J. Cheng, J. Liu, H. Kuang, and J. Wang. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. *IEEE Transactions on Medical Imaging*, 1(1):1–1, 2022.

[32] L. Yang, Y. Zhang, J. Chen, and et al. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 399–407, 2017.

[33] Z. Zhou, J. Shin, L. Zhang, and et al. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7340–7351, 2017.

[34] Z. Zhou, J. Y. Shin, S. R. Gurudu, and et al. Aft*: Integrating active learning and transfer learning to reduce annotation efforts. *arXiv preprint arXiv:1802.00912*, 2018.

[35] Z. Zhou, J. Shin, R. Feng, and et al. Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. *Journal of Digital Imaging*, 32(2):290–299, 2019.

[36] D. Yoo and I. S. Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019.

[37] A. Smailagic, P. Costa, A. Gaudio, and et al. O-medal: Online active deep learning for medical image analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):e1353, 2020.

[38] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A. González Ballester, and Gemma Piella. Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty. *Medical Image Analysis*, 75:102273, 2022.

[39] J. Yuan, X. Hou, Y. Xiao, and et al. Multi-criteria active deep learning for image classification. *Knowledge-Based Systems*, 172:86–94, 2019.

[40] S. Sinha, S. Ebrahimi, and T. Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

[41] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[42] Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran, Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, et al. Medal: Accurate and robust deep

active learning for medical image analysis. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 481–488. IEEE, 2018.

[43] K. He, X. Zhang, S. Ren, and et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[44] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015.