IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 February 2024, accepted 6 March 2024, date of publication 13 March 2024, date of current version 19 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376746

RESEARCH ARTICLE

Federated Active Learning (F-AL): An Efficient Annotation Strategy for Federated Learning

JIN-HYUN AHN, (Member, IEEE), YEEUN MA, SEOYUN PARK, AND CHEOLWOO YOU, (Member, IEEE)

Department of Information and Communication Engineering, Myongji University, Yongin, Gyeonggi-do 17058, Republic of Korea

Corresponding author: Cheolwoo You (cwyou@mju.ac.kr)

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00212836), and in part by 2022 Research Fund of Myongji University.

ABSTRACT Federated learning (FL) has been intensively investigated in terms of communication efficiency, privacy, and fairness. However, efficient annotation, which is a pain point in real-world FL applications, is less studied. In this project, we propose to apply active learning (AL) to the FL framework to reduce the annotation workload. We expect that the AL and FL can improve the performance of each other complementarily. In our proposed federated active learning (F-AL) method, the clients collaboratively execute the AL to obtain the instances which are considered informative to FL in a distributed optimization manner. We compare the test accuracies of the global FL models using the conventional random sampling strategy, client-level separate AL (S-AL), and the proposed F-AL. We empirically demonstrate that the F-AL outperforms baseline methods in image classification tasks.

INDEX TERMS Active learning, annotation, distributed learning, federated learning.

I. INTRODUCTION

Federated learning (FL) [1] enables the collaborative training from datasets residing on distributed clients with the help of a parameter server. In numerous previous works, including [2], [3], [4], and [5], the superiority of FL has been validated through numerical results and convergence analysis in independently identically distributed (IID) and non-IID datasets. While the recent literature related to FL primarily addresses communication efficiency, fairness, robustness, privacy of FL, and personalization, almost all of the previous works have assumed that the training datasets at clients are perfectly ready to be used for training.

However, the annotation step should not be overlooked or ignored for the practical implementations of FL, like other machine learning (ML), since the cost for labeling is generally high and might be even dominant over the FL itself. Considering this problem, we study the annotation strategies in the FL framework, where the clients participating in FL should label their datasets prior to FL execution. For the annotations, we apply active learning (AL) [6] at each client participating in FL. Because labeling all the instances is rarely a practical or cost-effective, AL aims to maximize the model's performance based on the fewest samples by selectively sampling and labeling the most informative instances.

To validate the proposed method, we establish a FL framework with the annotation step, where various active learning strategies in the FL are compared: 1) conventional FL with random sampling, 2) client-level separated active learning (S-AL), and 3) the proposed federated active learning (F-AL). In the F-AL, the clients collaboratively execute the AL to select the instances that are considered informative to FL in a distributed optimization manner. For the S-AL and F-AL, the state-of-the-art AL algorithms are incorporated.

The AL certainly outperforms the random sampling in the centralized learning. However, to the best of the authors' knowledge, there has been few work for considering the

The associate editor coordinating the review of this manuscript and approving it for publication was Fabrizio Marozzo.

AL in the FL framework and investigating the effect of AL on the performance of FL. The works [7], [8] empirically demonstrated that the S-AL outperforms the random selection (conventional FL) and AL is equally beneficial in both federated and centralized learning environments. This work also demonstrates that AL can surprisingly reduce the cost of labeling for and the proposed F-AL considerably improves the performance of AL in the FL environment. We summarize our contributions below:

- We establish a general FL framework combining with the annotation step. We evaluate the three types of methods: conventional FL with random sampling, S-AL, and F-AL. With the S-AL, the clients independently apply AL in their datasets. The F-AL encourages the clients' collaboration for AL.
- We empirically demonstrate that the AL is effective in the FL environment through various experiments with AL algorithms and datasets. The numerical result indicates that the AL methods outperform random sampling in terms of test accuracy of global FL models.
- We demonstrate that F-AL outperforms the other methods. We highlight that the F-AL magnifies the benefit of AL in the FL environment.

II. RELATED WORK & BACKGROUND

A. FEDERATED LEARNING

FL can be categorized into cross-device FL and cross-silo FL [9]. In both of FL, data is locally generated and stored while the data is centrally managed and distributed to clients in the setting of datacenter distributed learning. The cross-device FL supposes that the clients are an enormous number of mobile or IoT devices connected by Wi-Fi or slow connections. Therefore, uplink communication is the main bottleneck of performance. Furthermore, it generally encounters fresh training samples which are never seen before since most clients participate only once in an entire FL process.

On the other hand, cross-silo FL typically supposes that the distribution scale is 2-100 clients, which are generally different organizations or geo-distributed data centers such as hospitals or banks. Therefore, it supposes that all clients are available during the whole FL process, and the clients' datasets are repeatedly used for training from round to round. The performance degradation due to communication bottleneck is not as severe as the case of cross-device FL. Instead, the performance heavily depends on the number or quality of the training dataset [10].

FL can be executed in various ways in terms of optimization strategy of the knowledge among the clients. The most classic algorithms in FL are federated stochastic gradient descent (FedSGD), or federated averaging (FedAvg) [1] which are based on the averaging of the clients' parameters. Beyond the vanilla algorithms, FedProx [3] and FedDF [11] tackles the systems and statistical heterogeneity, FedMA [12] and FetchSGD [13] alleviate the communication bottleneck, and TERM [14] and Ditto [15] are related to the fairness and robustness in personalized FL.

B. ACTIVE LEARNING

AL selects the informative instances to be labeled prior to the other instances and aims to maximize the model's performance based on the fewest samples. It has been demonstrated that AL can considerably reduce the number of labeling samples and alleviate the heavy burden of cost for annotation [6], [16]. In fact, it has been proved that an effective AL strategy can theoretically obtain exponential acceleration in the efficiency of labeling [17]. Even when it is applied in the area of deep learning (DL), the cost saving in the annotation is much more fascinating since DL has its explicit limitation due to the high cost of labeling the numerous instances, even brutal in the professional field that requires rich knowledge [18], [19].

The sampling strategies of AL can be categorized into uncertainty-based sampling, representation-based sampling, other sampling strategies leverage the characteristic of deep learning such as learning loss (LL) [20], Monte-Carlo dropout (MC-dropout) [21], adversarial active learning [22], [23], and hybrid sampling using the strategies jointly. Uncertainty-based sampling [24], [25] queries the instances which are the most uncertain to the model trained on the current training samples. Representation-based sampling [26], [27] measures the representativeness of unlabeled samples and encourages the sampling strategy to select the instances from different areas of the distribution. Since the sampling strategy only concerned with uncertainty may skew the model due to the similarity of the sampled instances in a particular distribution, the balance between uncertainty and representativeness is one of the main issues in the performance of AL strategies [27].

Furthermore, most of the recent work related to AL focuses on the AL strategy for DL by leveraging the aspects of ML model such as estimated training loss [20], length of gradient [28] and MC dropout [21] for uncertainty estimation. Adversarial active learning [22], [23], [29], [30] trains a generative adversarial network (GAN) structured auxiliary network which learns a low dimensional latent space and discriminates the labeled and unlabeled samples in order to select unlabeled instances which are most different from the labeled instances. Furthermore, [31] recently proposed Maximum Classifier Discrepancy for Active Learning (MCDAL) which is the first work that leverages classifier discrepancy for sampling in active learning.

III. PROBLEM DEFINITION

This section provides FL framework where the annotation step is included before the execution of FL. We first introduce the FL environment comprising a parameter server and clients. The annotation step in the FL framework is described and formulated in more detail. Furthermore, we provide the mathematical modeling of AL.

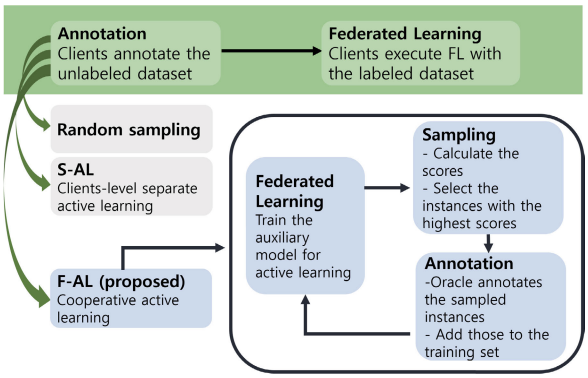

FIGURE 1. Annotation strategies for federated learning.

A. FEDERATED LEARNING ENVIRONMENT

We consider a cross-silo FL comprising a parameter server and M clients. The clients store their own local dataset $\mathcal{U}_m$, $m = 1, \ldots, M$ which are the unlabeled datasets. Before the start of FL, each m-th client selects the instances from $\mathcal{U}_m$ and labels the instances to obtain $\mathcal{D}_m = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ where $\mathcal{L}_m = \{\boldsymbol{x}_i\}$ is the selected instance from $\mathcal{U}_m$ and $\boldsymbol{y}_i$ is the label of $\boldsymbol{x}_i$. We denote the sampling function as $\mathcal{A}(\cdot)$ and the selected instances, $\mathcal{L}_m$, as

$$\mathcal{L}_m = \mathcal{A}(\mathcal{U}_m), \tag{1}$$

for $m = 1, \ldots, M$.

Let $\boldsymbol{\theta} \in \mathbb{R}^D$ denote the global model to be optimized in FL. The local loss $F_m(\boldsymbol{\theta})$ at the m-th client is $F_m(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_m|}\sum_{\boldsymbol{u}\in\mathcal{D}_m} f(\boldsymbol{\theta}, \boldsymbol{u})$, where $\mathcal{D}_m$ is the labeled dataset at the m-th client and $f(\cdot)$ is the loss function determined by the network model. Accordingly, a global loss $F(\boldsymbol{\theta})$ can be defined as $F(\boldsymbol{\theta}) = \frac{1}{\left|\bigcup_{m=1}^{M}\mathcal{D}_m\right|}\sum_{\boldsymbol{u}\in\bigcup_{m=1}^{M}\mathcal{D}_m} f(\boldsymbol{\theta}, \boldsymbol{u})$. The goal of FL is to find the optimized parameter $\boldsymbol{\theta}^*$ minimizing the global loss, namely $\boldsymbol{\theta}^* = \underset{\theta}{\operatorname{argmin}}\, F(\boldsymbol{\theta})$, without sharing the local datasets.

For this, FedSGD [1] utilizes the iterative stochastic gradient descent (SGD), allowing the parallel computation of gradients among the clients. The parameter vector $\boldsymbol{\theta}_t$ at the t-th iteration is updated according to $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \sum_{m=1}^{M} \frac{n_m}{n} \boldsymbol{g}_m(\boldsymbol{\theta}_t)$, where η_t is the learning rate at the t-th iteration, $n = \sum_{m=1}^{M} n_m$, $n_m = |\mathcal{D}_m|$ and $\boldsymbol{g}_m(\boldsymbol{\theta}_t) \in \mathbb{R}^D$ is the stochastic gradient of $\boldsymbol{\theta}_t$ computed at the m-th client as $\boldsymbol{g}_m(\boldsymbol{\theta}_t) = \frac{1}{|\mathcal{D}_m|}\sum_{\boldsymbol{u}\in\mathcal{D}_m} \nabla f(\boldsymbol{\theta}_t, \boldsymbol{u})$. The update is equivalently given by

$$\boldsymbol{\theta}_{t+1} = \sum_{m=1}^{M} \frac{n_m}{n} \boldsymbol{\theta}_{t+1}^m, \tag{2}$$

where

$$\boldsymbol{\theta}_{t+1}^m = \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_m(\boldsymbol{\theta}_t). \tag{3}$$

As a consequence, $\boldsymbol{\theta}^*$ can be achieved without collecting the local datasets in FedSGD, since it is sufficient to exchange only the gradients, $\{\boldsymbol{g}_m(\boldsymbol{\theta}_t)\}_{m=1}^{M}$. The FL update of θ is described based on the FedSGD here but it can be easily extended to utilize the other optimization algorithms. The overall FL frameworks are summarized in Algorithm 1.

Algorithm 1 Federated Learning With Annotation Step

Input: unlabeled datasets, $\{\mathcal{U}_m\}_{m=1}^{M}$
initialized model, $\boldsymbol{\theta}_1$
learning rate, $\{\eta_t\}_t$

Annotation step:
for $m = 1$ **to** M **do**
  annotate $\mathcal{L}_m = \mathcal{A}(\mathcal{U}_m)$ to obtain $\mathcal{D}_m$
end for
FL step:
for each round $t = 1$ **to** T **do**
  m-**th client executes:**
  do multiple iterations of (3)
  send $\boldsymbol{\theta}_{t+1}^m$ to the server
  Server executes:
  average model parameters as in (2)
  send $\boldsymbol{\theta}_{t+1}$ to the clients
end for
return $\boldsymbol{\theta}_{T+1} = \texttt{FedAvg}\left(\{\mathcal{D}_m\}_m^M \mid \boldsymbol{\theta}_1, \{\eta_t\}_t\right)$

B. ACTIVE LEARNING

In the proposed FL framework, we introduce the sampling function, $\mathcal{A}(\cdot)$, which finds the instances to be labeled from the unlabeled dataset prior to the process of FL. For an example of random sampling, the acquired sample instances from the unlabeled dataset $\mathcal{U}$ is $\mathcal{A}(\mathcal{U}) = \texttt{random}(\mathcal{U}, b)$, where $\texttt{random}(\mathcal{U}, b)$ is to randomly choose the b instances from $\mathcal{U}$. In terms of the sampling function, the goal of AL is to find the best sampling function which selects the most informative and effective instances to the performance of the main task.

Most of the AL algorithms generally searches instances with the highest score in the unlabeled data pool [32] as

$$\mathcal{A}(\mathcal{U}) = \underset{\mathcal{L}\subseteq\mathcal{U},\ |\mathcal{L}|=b,\ x\in\mathcal{L}}{\operatorname{argmax}} S(x), \tag{4}$$

where b is the budget of sampling, and $S(x)$ is the score function of x. The score function of effective AL should perfectly reflect the potential informativeness of instances in the unlabeled dataset. Hence, the AL models can be described by how to design $S(\cdot)$. The score includes uncertainty, diversity, density, training loss, and dissimilarity to the labeled dataset.

Since the informativeness depends on the current labeled dataset, the score function is also conditioned on the current state of the labeled dataset. Hence, the score is generally calculated based on the trained model with the current labeled dataset, namely

$$S(x) = S(x \mid \mathcal{D}) = S(x \mid \boldsymbol{\phi}(\mathcal{D})), \tag{5}$$

where $\boldsymbol{\phi}(\mathcal{D})$ is the auxiliary model that is trained with the labeled dataset $\mathcal{D}$, starting from the randomly initialized

Algorithm 2 Active Learning, $\mathcal{A}(\cdot)$

Input: unlabeled dataset, $\mathcal{U}^1$
initially labeled dataset, $\mathcal{D}^1$
number of AL round, K
initialized models for $\boldsymbol{\phi}$, $\{\boldsymbol{\phi}^k\}_{k=1}^K$
number of annotation budget, b

for $k = 1$ **to** K **do**
  train $\boldsymbol{\phi}\left(\mathcal{D}^k\right)$, starting from $\boldsymbol{\phi}^k$
  sample $\mathcal{L}^k$,

$$\mathcal{L}^k = \underset{\mathcal{L}\subseteq\mathcal{U}^k,\ |\mathcal{L}|=\frac{b}{K},\ x\in\mathcal{L}}{\operatorname{argmax}} S\left(x \,\middle|\, \boldsymbol{\phi}\left(\mathcal{D}^k\right)\right) \tag{6}$$

  $\hat{\mathcal{D}}^k = \texttt{annotate}\left(\mathcal{L}^k\right)$
  $\mathcal{D}^{k+1} = \mathcal{D}^k \cup \hat{\mathcal{D}}^k$
  $\mathcal{U}^{k+1} = \mathcal{U}^k - \mathcal{L}^k$
end for
return $\mathcal{D}^{K+1}$ with size of $\left|\mathcal{D}^1\right| + b$

Algorithm 3 Federated Active Learning, $\mathcal{A}(\cdot)$

Input: unlabeled dataset, $\left\{\mathcal{U}_m^1\right\}_{m=1}^M$
initially labeled dataset, $\left\{\mathcal{D}_m^1\right\}_{m=1}^M$
number of AL round, K
initialized models for $\boldsymbol{\phi}$, $\{\boldsymbol{\phi}^k\}_{k=1}^K$
number of annotation budget, $\{b_m\}_{m=1}^M$

for $k = 1$ **to** K **do**
  FL step:
  train $\boldsymbol{\phi}_{FL}^k$, starting from $\boldsymbol{\phi}^k$

$$\boldsymbol{\phi}_{FL}^k = \texttt{FedAvg}\left(\left\{\mathcal{D}_m^k\right\}_{m=1}^M \,\middle|\, \boldsymbol{\phi}^k, \left\{\eta_t^k\right\}_t\right) \tag{7}$$

  Sampling step:
  for $m = 1$ **to** M **do**
    sample $\mathcal{L}_m^k$,

$$\mathcal{L}_m^k = \underset{\mathcal{L}\subseteq\mathcal{U}_m^k,\ |\mathcal{L}|=\frac{b_m}{K},\ x\in\mathcal{L}}{\operatorname{argmax}} S\left(x \,\middle|\, \boldsymbol{\phi}_{FL}^k\right) \tag{8}$$

    $\hat{\mathcal{D}}_m^k = \texttt{annotate}\left(\mathcal{L}_m^k\right)$
    $\mathcal{D}_m^{k+1} = \mathcal{D}_m^k \cup \hat{\mathcal{D}}_m^k$
    $\mathcal{U}_m^{k+1} = \mathcal{U}_m^k - \mathcal{L}_m^k$
  end for
end for
return $\mathcal{D}_m^{K+1}$ with size of $\left|\mathcal{D}_m^1\right| + b_m, m = 1, \ldots, M$

model. Furthermore, AL adopts multiple rounds for sampling and gradually samples from the unlabeled dataset. When it is desired to add b instances to be labeled after K rounds, it samples b/K instances at each round. We summarize the description of AL model, $\mathcal{A}(\cdot)$, in Algorithm 2.

IV. FEDERATED ACTIVE LEARNING (F-AL)

This section introduces AL methods in the FL framework: S-AL and F-AL. In the benchmark scheme of conventional FL adopting random sampling, we set $\mathcal{A}(\mathcal{U}) = \texttt{random}(\mathcal{U}, b)$ in the Algorithm 1.

A. SEPARATE ACTIVE LEARNING (S-AL)

In S-AL, the clients separately perform the AL before the FL execution. With S-AL, the m-th client applies $\mathcal{A}(\cdot)$ of Algorithm 2 to its unlabeled dataset at the annotation step in the FL framework. The S-AL directly leverages the AL in the FL framework, including the annotation step.

B. FEDERATED ACTIVE LEARNING (F-AL)

In S-AL, the clients independently accomplish AL and achieve the instances which are informative to the local datasets as in (6). At the k-th round, the m-th client selects x with the highest score, $S\left(x \,\middle|\, \boldsymbol{\phi}\left(\mathcal{D}_m^k\right)\right)$, where $\mathcal{D}_m^k$ denotes $\mathcal{D}^k$ in the Algorithm 2 in the perspective of m-th client.

Since the clients execute FL after the annotation step, however, it should be the main objective to obtain instances which are informative to the aggregate labeled dataset, $\mathcal{D}_{total}^k = \bigcup_{m=1}^M \mathcal{D}_m^k$ as in (5). Therefore, the score function in F-AL is conditioned on $\mathcal{D}_{total}^k$ and defined as $S\left(x \,\middle|\, \boldsymbol{\phi}\left(\mathcal{D}_{total}^k\right)\right)$. But, $\boldsymbol{\phi}\left(\mathcal{D}_{total}^k\right)$ cannot be built because $\mathcal{D}_m^k$, $m = 1, \ldots, M$ should not be compiled to satisfy the constraint of FL. Thus, we replace $\boldsymbol{\phi}\left(\mathcal{D}_{total}^k\right)$ with the model trained by FL, $\boldsymbol{\phi}_{FL}^k$, which is

$$\boldsymbol{\phi}_{FL}^k = \texttt{FedAvg}\left(\left\{\mathcal{D}_m^k\right\}_{m=1}^M \,\middle|\, \boldsymbol{\phi}^k, \left\{\eta_t^k\right\}_t\right). \tag{9}$$

Accordingly, with F-AL, clients carry out FL to obtain the score function that represents the informativeness of the aggregate labeled dataset.

As a more clear perspective for the explanation, uncertainty [24], [25] can illustrate the ground why $\boldsymbol{\phi}_{FL}^k$ should be leveraged for the calculation of score function in order to improve FL performance. If the AL applies uncertainty-based sampling or the sampling related to uncertainty, referred as to task aware AL in [23], it utilizes the uncertainty score, which is measured by the main task model trained with the current labeled dataset. Therefore, the auxiliary model is the main task model, namely, $\boldsymbol{\phi}\left(\mathcal{D}^k\right) = \boldsymbol{\theta}\left(\mathcal{D}^k\right)$ in Algorithm 2 or the set of auxiliary models includes the main tasks model in the case of several auxiliary models. Hence, we remark that the auxiliary model should also be obtained through FL since the main task model is trained by FL.

After attaining $\boldsymbol{\phi}_{FL}^k$ in F-AL, the instances can be ideally sampled as

$$\mathcal{L}^k = \underset{\mathcal{L}\subseteq\mathcal{U}^k,\ |\mathcal{L}|=\frac{b}{K},\ x\in\mathcal{L}}{\operatorname{argmax}} S\left(x \,\middle|\, \boldsymbol{\phi}_{FL}^k\right), \tag{10}$$

where $\mathcal{U}^k = \bigcup_{m=1}^M \mathcal{U}_m^k$, $b = \sum_{m=1}^M b_m$. Under the annotation workload condition that the m-th client annotates b_m/K

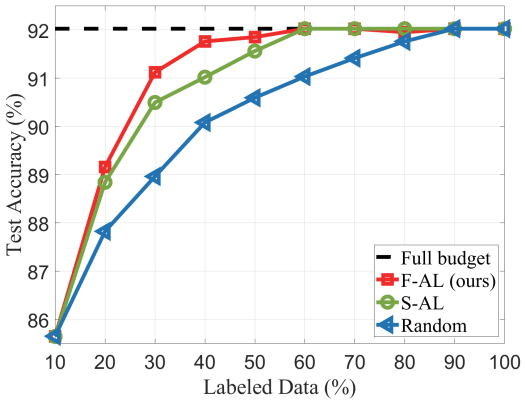

FIGURE 2. Test accuracies of global model trained by FL per rounds on Fashion-MNIST (MCDAL).

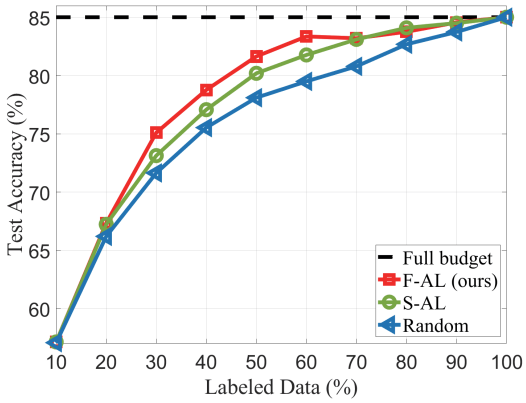

FIGURE 3. Test accuracies of global model trained by FL per rounds on CIFAR-10 (MCDAL).

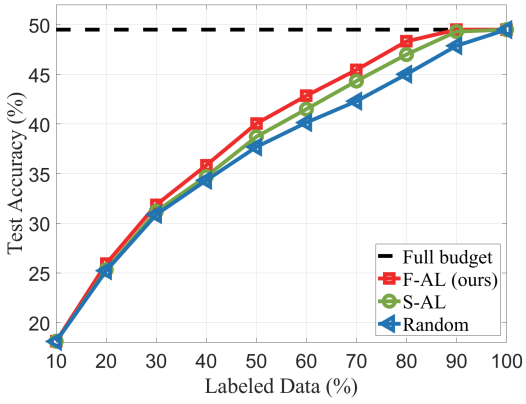

FIGURE 4. Test accuracies of global model trained by FL per rounds on CIFAR-100 (MCDAL).

instances at each round, we have $\mathcal{L}^k = \bigcup_{m=1}^{M} \mathcal{L}_m^k$, where

$$\mathcal{L}_m^k = \underset{\mathcal{L}\subseteq\mathcal{U}_m^k,\ |\mathcal{L}|=\frac{b_m}{K},\ x\in\mathcal{L}}{\operatorname{argmax}} S\left(x \,\middle|\, \boldsymbol{\phi}_{FL}^k\right). \tag{11}$$

Therefore, each m-th client samples $\mathcal{L}_m^k$ as in (11) and follows the remaining steps in Algorithm 2. In fact, the sampling step in (10) can be executed at the server by exchanging the scores and indices of instances. However, we do not go any further since it might break the fairness of annotation workload among clients.

V. EXPERIMENTS

This section provides the implementation details and the numerical results with related discussion. We compare the performance of FL using the random sampling, S-AL, and the proposed F-AL in image classification tasks. The annotation strategies are applied for the annotation step in the Algorithm 1, where the test accuracy of the obtained model is measured for the performance metric. For the image classification tasks, we evaluate the performances of the annotation strategies on the classical public datasets, Fashion-MNIST [33], CIFAR-10 [34], and CIFAR-100 [34]. The Fashion-MNIST dataset is a more challenging alternative dataset for the MNIST dataset. It consists of a training dataset of 60,000 images for 10 types of clothing and a test dataset of 10,000 images. CIFAR-10 and CIFAR-100 contain 50,000 training images and 10,000 test images. CIFAR-10 has 10 classes, while CIFAR-100 has 100 classes.

A. ACTIVE LEARNING ALGORITHMS

First, we evaluate the performance of annotation strategies when the AL model is the recently proposed Maximum Classifier Discrepancy for Active Learning (MCDAL) [31] which is one of the state-of-the-art AL algorithms. It utilizes the prediction discrepancies between two auxiliary classifiers after learning the auxiliary classifiers to maximize the discrepancies. It replaces the classic uncertainty with the discrepancies in the predictions of the auxiliary classifiers. Accordingly, the main task model and two auxiliary classifiers are obtained by a federate manner as in (8).

For more discussion, we evaluate the performance of annotation strategies for the various kinds of AL models to achieve consistency in performance comparison. The AL models includes the classic uncertainty-based sampling with maximum entropy [24], MC-dropout with maximum entropy [21], Learning Loss (LL) [20], and MCDAL [31]. All of the algorithms consider the main task model as the auxiliary model for AL. In the LL, the learning loss module is also included in the auxiliary models.

B. IMPLEMENTATION DETAILS

In the experiments, we assume that $M = 5$ clients respectively have disjoint 10000 images where 10% of the dataset is initially labeled. In our active learning setup, the 10% of the dataset is added to the labeled dataset at the sampling step of each round. We repeat this AL rounds until the total dataset is labeled. Hence, we set $b = 10000$, $K = 10$, and measure the test accuracy of FL model at each k-th round of AL, $\texttt{FedAvg}\left(\left\{\mathcal{D}_m^k\right\}_m^M \,\middle|\, \boldsymbol{\theta}_1^k, \left\{\eta_t^k\right\}_t\right)$.

We apply the Resnet-18 [35] for the base architecture of main task model for all the exemplary tasks. In the FL

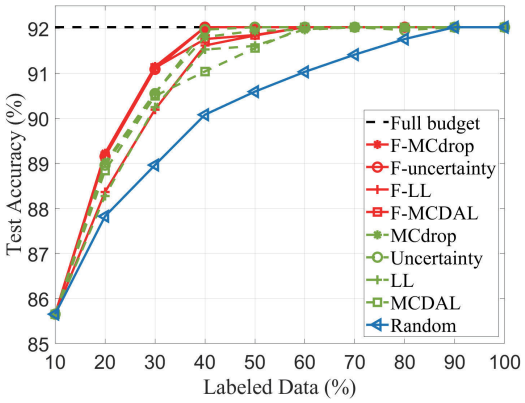

FIGURE 5. Test accuracies of global model trained by FL per rounds on Fashion-MNIST.

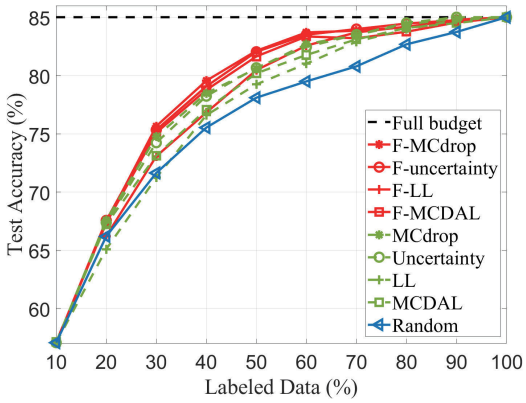

FIGURE 6. Test accuracies of global model trained by FL per rounds on CIFAR-10.

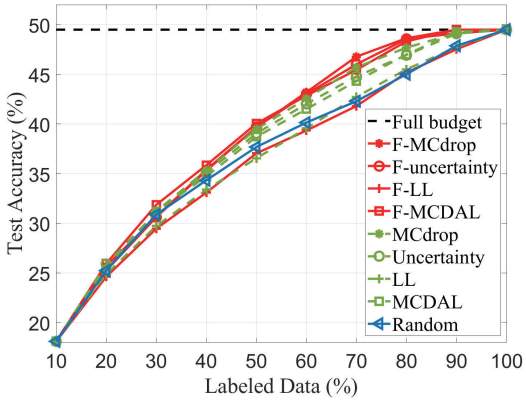

FIGURE 7. Test accuracies of global model trained by FL per rounds on CIFAR-100.

implementation, the main task models are optimized by SGD with the learning rate of 5×10^{-2} and learning rate decay of 0.997 per global iteration. The number of the local epoch is 1, and the global iteration ends when the training loss at the clients decrease below thresholds, 1×10^{-3}, 5×10^{-4}, and 1.5×10^{-3} for Fashion-MNIST, CIFAR-10, and CIFAR-100, respectively. In the independent learning for S-AL, we use SGD with a learning rate 1×10^{-2} and step decay of 0.997 at every epoch. Independent learning follows the same stopping criteria as FL. We use random horizontal flips for data augmentations. For the result of the experiments, we use the average accuracy of three runs.

C. PERFORMANCE COMPARISON

Fig. 2-4 illustrate the performance of random sampling (conventional FL), S-AL (benchmark), and F-AL (ours) which are the annotation strategies for FL. The AL model is MCDAL, and the datasets are Fashion-MNIST, CIFAR-10, and CIFAR-100. Full budget in the figures denotes the performance of FL when all the clients have 100% labeled dataset. On the Fashion-MNIST, F-AL and S-AL considerably outperform random sampling, and the proposed F-AL shows the best performance compared to the other strategies. In particular, the average improvement compared to random sampling is 1.1% and 1.6% for S-AL and F-AL, respectively, at the 2nd round and 3rd round, before converging to the performance of the full budget.

On the CIFAR-10, F-AL and S-AL outperform random sampling, and the proposed F-AL shows the best performance compared to the other strategies, same as the case of Fashion-MNIST. The average improvement compared to random sampling is 1.6% and 2.4% for S-AL and F-AL, respectively, before the 10th round. At the half of the rounds, the improvement is 2.3% and 3.9% for S-AL and F-AL, respectively when the performance of random sampling is 79.5%.

On the CIFAR-100, it is also observed that the F-AL and S-AL show better performance than the performance of random sampling. The average improvement compared to random sampling is 1.1% and 2.0% for S-AL and F-AL, respectively, before the 10th round and the improvements are 2.0% and 3.2% at the 7th round while the test accuracy of random sampling is 42.3%. Fig. 2-4 demonstrate that the proposed F-AL outperforms the baseline methods in the image classification of Fashion-MNIST, CIFAR-10, and CIFAR-100.

D. EXTENDED RESULTS FOR VARIOUS AL MODELS

In order to demonstrate that our proposed F-AL outperforms the baseline methods for the general AL models, we extend the experiment with MCDAL in Fig. 5-7. We consider uncertainty-based sampling, MC-dropout with maximum entropy, LL, and MCDAL. Fig. 5-7 illustrate that F-AL outperforms S-AL and random sapling for the considered AL models. The only conflicting case is when LL is applied on the CIFAR-100, as observed in Fig. 7. In Table 1, we illustrate the performance of the algorithms in the case of CIFAR-10.

E. DISCUSSION

In Fig. 5-7 and Table 1, it was first observed that uncertainty-based sampling and MC-dropout, which directly utilizes the uncertainty, show the best performance across most rounds of AL, and they have the largest performance increase

TABLE 1. Test accuracies of global model trained by FL per rounds on CIFAR-10.

methods	Labeled Data (%) 10%	20%	30%	40%	50%	60%	70%	80%	90%
F-MCdrop	57.07	67.41	**75.66**	**79.56**	**82.13**	**83.71**	**83.85**	84.18	**84.80**
F-uncertainty	57.07	**67.57**	**75.34**	**79.08**	**82.04**	**83.51**	**84.00**	**84.47**	84.76
F-LL	57.07	66.21	73.07	76.83	80.47	82.58	**83.71**	84.20	84.70
F-MCDAL	57.07	**67.88**	**75.26**	**78.56**	**81.53**	**83.24**	83.34	84.04	84.79
MCdrop	57.07	**67.59**	74.77	78.49	80.61	82.70	83.50	**84.41**	**84.88**
Uncertainty	57.07	67.38	74.24	78.24	80.67	82.46	83.52	**84.53**	**85.01**
LL	57.07	65.09	71.24	76.60	79.25	81.02	82.91	83.98	84.76
MCDAL	57.07	67.09	73.13	77.84	80.00	81.86	83.15	83.79	84.45
Random	57.07	66.20	71.64	75.54	78.10	79.50	80.79	82.69	83.75

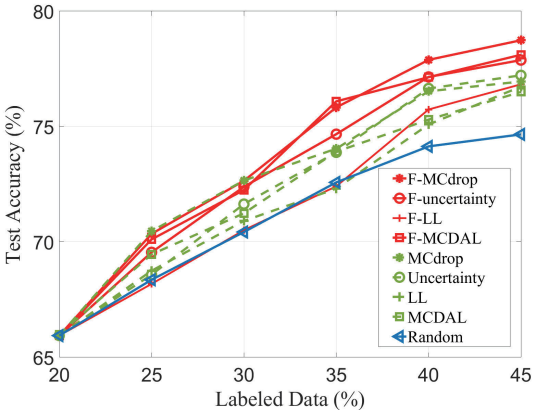

FIGURE 8. Test accuracies of global model trained by FL per round on CIFAR-10 (heterogeneous).

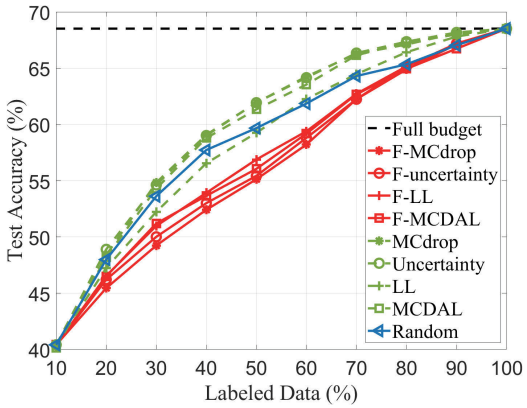

FIGURE 9. Test accuracies of model trained by independent learning per rounds on CIFAR-10.

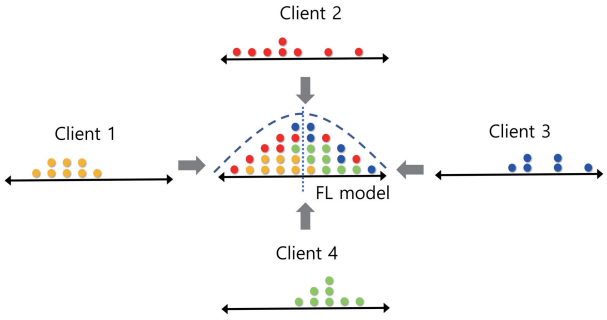

FIGURE 10. Distributions of the sampled instances.

by F-AL. In the previous literature [20], [31], it is validated that the LL and MCDAL outperform the classic uncertainty-based sampling and MC-dropout, contrary to the results in our experiments. In fact, LL and MCDAL learn the classifiers for discrepancy and the loss prediction module, respectively, in addition to the main task model, using the *unlabeled dataset*. Compared to the large-scale dataset stored at one client in the literature [20], [31], multiple clients relatively have a much less number of instances in the unlabeled dataset for the distributed setting, e.g., 5 clients respectively have 20% of the total dataset in our experiments. This insufficiency of the unlabeled dataset in the FL environment causes the comparable performance of LL and MCDAL with the classical uncertainty-based AL models.

1) HETEROGENEOUS DATASET

In Fig. 2-7, we provide the numerical comparison when the distribution of dataset is homogeneous among the clients. Data heterogeneity is a common feature for both cross-silo and cross-device federated learning. Furthermore, the AL is highly relevant to the data distribution so that investigating the effect of heterogeneity is valuable. We are extending the numerical result to the case of heterogeneous distribution among the clients. For the case of CIFAR-10, we provide the numerical comparison when the distributions include the imbalance among the labels of instances. Each client has two target labels and has 2500 unlabeled instances for each target label. And 5000 instances are drawn from the other labels. In Fig. 8, it is also observed that F-AL and S-AL outperform the random sampling and the proposed F-AL shows the best performance compared to the other strategies, in a same manner with the case of homogeneous data distribution.

2) PERFORMANCE OF INDEPENDENT LEARNING

Through Fig. 2-7 and Table 1, it has been demonstrated that AL is effective in FL environment, and the proposed

F-AL outperforms the conventional random sampling and S-AL. For more discussion, we investigate the effect of F-AL in the perspective of local dataset. For this, each client solely trains the main task model with the local dataset after achieving the labeled dataset via the AL strategies. Fig. 9 illustrates the average test accuracy of the models trained at the clients on CIFAR-10. It is observed that F-AL considerably decreases the performance of IL. In contrast, the S-AL certainly outperforms random sampling since S-AL samples the informative instances to the current local dataset. With F-AL, the clients collaborate to sample the informative instances to the aggregate datasets, not the local dataset. It becomes a solid constraint to the sampling of clients in the perspective of local datasets since each client with F-AL does not sample the instances that are not informative to the aggregate dataset even though the instances are informative to its datasets. As illustrated in Fig. 10, the aggregate dataset, which is sampled by F-AL, performs excellent for FL even though the sampled instances can be biased at the distribution of local datasets.

VI. CONCLUSION

In this paper, we focused on the active learning (AL) and sampling strategies into the FL framework to reduce the annotation workload. In our proposed federated active learning (F-AL) method, the clients collaboratively perform the AL to obtain the instances that can maximally improve the global model of FL. We empirically demonstrate that F-AL outperforms conventional random sampling strategy, client-level separate AL (S-AL) for the various AL models on the image classification applications such as Fashion-MNIST, CIFAR-10, and CIFAR-100.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[2] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," 2017, *arXiv:1705.10467*.
[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.
[4] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2019.
[5] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," 2019, *arXiv:1907.02189*.
[6] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. TR-2009, 2009.
[7] L. Ahmed, K. Ahmad, N. Said, B. Qolomany, J. Qadir, and A. Al-Fuqaha, "Active learning based federated learning for waste and natural disaster image classification," *IEEE Access*, vol. 8, pp. 208518–208531, 2020.
[8] N. Aussel, S. Chabridon, and Y. Petetin, "Combining federated and active learning for communication-efficient distributed failure prediction in aeronautics," 2020, *arXiv:2001.07504*.
[9] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, and R. Cummings, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.
[10] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, and E. Herrera-Viedma, "Data set quality in machine learning: Consistency measure based on group decision making," *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107366.
[11] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," 2020, *arXiv:2006.07242*.
[12] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," 2020, *arXiv:2002.06440*.
[13] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8253–8265.
[14] T. Li, A. Beirami, M. Sanjabi, and V. Smith, "Tilted empirical risk minimization," 2020, *arXiv:2007.01162*.
[15] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6357–6368.
[16] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, Oct. 2021.
[17] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," *J. Comput. Syst. Sci.*, vol. 75, no. 1, pp. 78–89, Jan. 2009.
[18] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2007, vol. 19, no. 1, pp. 153–160.
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[20] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 93–102.
[21] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
[22] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5971–5980.
[23] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8162–8171.
[24] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. SIGIR*. New York, NY, USA: Springer, 1994, pp. 3–12.
[25] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9368–9377.
[26] Y. Geifman and R. El-Yaniv, "Deep active learning over the long tail," 2017, *arXiv:1711.00941*.
[27] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017, *arXiv:1708.00489*.
[28] A. Freytag, E. Rodner, and J. Denzler, "Selecting influential examples: Active learning with expected model output changes," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2014, pp. 562–577.
[29] S. Wang, Y. Li, K. Ma, R. Ma, H. Guan, and Y. Zheng, "Dual adversarial network for deep active learning," in *Computer Vision—ECCV*, Glasgow, U.K. Springer, 2020, pp. 680–696.
[30] B. Zhang, L. Li, S. Yang, S. Wang, Z.-J. Zha, and Q. Huang, "State-relabeling adversarial active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8756–8765.
[31] J. Won Cho, D.-J. Kim, Y. Jung, and I. So Kweon, "MCDAL: Maximum classifier discrepancy for active learning," 2021, *arXiv:2107.11049*.
[32] A. K. McCallumzy and K. Nigamy, "Employing em and pool-based active learning for text classification," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 1998, pp. 359–367.
[33] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[34] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

JIN-HYUN AHN (Member, IEEE) received the B.S. and M.S. degrees in mathematics and the Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2013, 2016, and 2020, respectively. From January to September 2019, he was a Visiting Researcher with the King's Communications, Learning and Information Processing Laboratory, King's College London, U.K. He was a Research Associate with KAIST, from September 2019 to September 2020. He was a Research Associate with the MGH/BWH Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, from October 2021 to August 2022. He is currently an Assistant Professor with the Department of Information and Communication Engineering, Myongji University, Yongin, Gyeonggi-do, South Korea. His research interests include probability theory, communication theory, and machine learning, with a specific focus on federated learning. He served as a Reviewer for many journals, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

YEEUN MA received the B.S. degree in mathematics from Myongji University, Yongin, Gyeonggi-do, South Korea, in 2024. She is currently pursuing the M.S. degree in information and communication engineering with Myongji University, under the supervision of Prof. Jin-Hyun Ahn. Her research interests include federated learning, reinforcement learning, and machine learning, with a specific focus on federated learning.

SEOYUN PARK received the B.S. degree in information and communication engineering from Myongji University, Yongin, Gyeonggi-do, South Korea, in 2024, where she is currently pursuing the M.S. degree in information and communication engineering. Her research interests include reinforcement learning, computer vision, and machine learning, with a specific focus on reinforcement learning.

CHEOLWOO YOU (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics engineering from Yonsei University, Seoul, Republic of Korea, in 1993, 1995, and 1999, respectively. From January 1999 to April 2003, he was a Senior Research Engineer with LG Electronics, Gyeonggi, South Korea. From 2003 to 2004, he was a Senior Research Engineer with EoNex, Songnam, South Korea. From August 2004 to July 2006, he was with Samsung Electronics, Suwon, South Korea. Since September 2006, he has been with the Department of Information and Communications Engineering, Myongji University, Gyeonggi, Yongin. His research interests include next generation communication systems, artificial intelligence, air I/F technologies in the international standards, communication theory, and signal processing. He is currently interested in the 5G/6G communication systems, machine and deep learning, AR/VR, IoE/AIoT, and V2X.