

KnFu: Effective Knowledge Fusion

S. Jamal Seyedmohammadi*, S. Kawa Atapour[†], Jamshid Abouei[†], Arash Mohammadi*

* Concordia Institute of Information Systems Engineering (CIISE), Concordia University, Montreal, Canada

[†] Dept. of Electrical Engineering, Yazd University, Yazd, Iran

Abstract—Federated Learning (FL) is a decentralized approach that allows for collaborative training of Machine Learning (ML) models across multiple local nodes, ensuring data privacy and security while leveraging diverse datasets. Conventional FL, however, is susceptible to gradient inversion attacks, restrictively enforces a uniform architecture on local models, and suffers from model heterogeneity (model drift) due to non-IID local datasets. To mitigate some of these challenges, the new paradigm of Federated Knowledge Distillation (FKD) has emerged. FKD is developed based on the concept of Knowledge Distillation (KD), which involves extraction and transfer of a large and well-trained teacher model’s knowledge to lightweight student models. FKD, however, still faces the model drift issue. Intuitively speaking, not all knowledge is universally beneficial due to the inherent diversity of data among local nodes. This calls for innovative mechanisms to evaluate the relevance and effectiveness of each client’s knowledge for others, to prevent propagation of adverse knowledge. In this context, the paper proposes Effective Knowledge Fusion (KnFu) algorithm that evaluates knowledge of local models to only fuse semantic neighbors’ effective knowledge for each client. The KnFu is a personalized effective knowledge fusion scheme for each client, that analyzes effectiveness of different local models’ knowledge prior to the aggregation phase. In this context, closeness of clients’ knowledge is measured by estimating the class distributions of local datasets based on the transmitted localized knowledge. Comprehensive experiments were performed on MNIST and CIFAR10 datasets. KnFu outperforms its FL-based, FKD-based, and local training baselines in scenarios where the clients have small datasets with intermediate degree of heterogeneity. KnFu’s source code is accessible through the following link: <https://github.com/jamal94sm/KnFu-Effective-Knowledge-Fusion.git>.

Index Terms—Personalized Federated Learning, Clustered Knowledge Distillation, Selective Knowledge Distillation

I. INTRODUCTION

FEDERATED Learning (FL) has recently gained considerable attention, as an alternative to the centralized learning paradigm, in various domains including but not limited to computer vision, healthcare [1], and natural language processing [2]. Generally speaking, FL resolves some practical challenges of centralized learning frameworks such as users’ privacy issues and the communication cost of transmitting raw data from users/silos to the Fusion Centre (FC). Conventional FL methods aim to collaboratively train a global model by aggregating parameters of the clients’ models without sharing their private data [3]. Such methods, however, pose the following new challenging problems: (i) Privacy concerns arising

from gradient inversion attacks; (ii) Communication overhead of iterative transmission/reception of model parameters; (iii) Enforcing a uniform model architecture on clients, and; (iv) Model heterogeneity (model drift) resulted from non-IID local datasets [4], [5].

To mitigate some of the above mentioned challenges of conventional FL solutions, the new paradigm of Federated Knowledge Distillation (FKD) [6], [7] has been introduced that integrates the concept of Knowledge Distillation (KD) with FL. KD involves extracting the knowledge of a large and well-trained teacher model and transferring it to a lightweight student model by mimicking the teacher’s predictions on a transfer set. In FKD, clients share only their local knowledge, i.e., predictions on the transfer set, with the server rather than their local model parameters. This leads to a more privacy-preserving framework, reduced communication overhead, and allowing heterogeneous model architectures among clients. While FKD presents effective advantages to resolve conventional FL’s problems, it poses some new difficulties, including: (i) Requiring a transfer set to extract local knowledge of clients, and; (ii) Imposing computation overhead on local devices. Additionally, the model drift issue still remains as an open challenge. Since clients hold non-IID local datasets, the models trained locally would be heterogeneous, resulting in non-IID local knowledge among clients. Therefore, aggregating the local knowledge of a specific client with that of other clients may lead to adverse impacts on the client’s local model resulting in significant performance degradation. Consequently, there has been a surge of recent interest devising innovative solutions to alleviate these issues [8]–[14]. This field, however, is still in its infancy. The paper aims to further advance the research in this domain.

Related Works: In [8] an adaptive KD approach is proposed, inspired by multitask learning methods, to adaptively adjust the weight of different distillation paths of an ensemble of teachers. Such an approach prevents negative impacts of some paths on the generalization performance of the student models. Reference [9] studied whether all or partial knowledge of a model is effective. A generic knowledge selection method is presented to select and distill only certain knowledge by either fixing the knowledge selection threshold or changing it progressively during the training process as the teacher’s confidence is enhanced. A selective knowledge-sharing mechanism is proposed in [10] to address the misleading and ambiguous knowledge fusion challenge resulting from non-IID

local datasets and absence of a well-trained teacher model. The client-side selector chooses accurate predictions that match the ground-truth labels. Meanwhile, the server-side selector identifies the precise prediction by their entropy values. Precise knowledge has low entropy, while ambiguous predictions have high entropy and uncertainty.

Reference [11] analyzed the effect of local predicted logits on the convergence rate. To improve the convergence rate, a knowledge selection method is proposed to schedule the predicted logits for efficient knowledge aggregation. In addition, a threshold-based approach is presented to optimize the local model updating options with/without knowledge distillation for each edge device to reduce the performance degradation of local models resulted from ambiguous knowledge. The COMET approach is proposed in [12] introducing the clustered knowledge distillation concept, i.e., forming localized clusters from clients with similar data distribution. Each client then uses the aggregated knowledge of its cluster rather than following the average logits of all clients. Such an approach prevents performance degradation by learning from clients with considerably different data distributions. In the local updating phase of each client, the loss function comprises a cross-entropy function along with a regularization term, which is an l_2 -norm between the local and average predictions of the corresponding cluster.

KT-pFL [13] proposed a personalized group knowledge distillation algorithm, updating the personalized soft prediction of each client through a linear combination of all local predictions by a knowledge coefficient matrix. This matrix adaptively adjusts the collaboration among clients with similar data distribution and is parameterized to be trained simultaneously with the models. MetaFed [14] presents a trustworthy personalized FL that achieves a personalized model for each federation without a central server using cyclic knowledge distillation. Its training process is split into two parts: common knowledge accumulation and personalization. In the first part, it leverages the validation accuracy on the current federation's validation data to decide whether to completely keep the previous federation's knowledge and fine-tune it or just use it to update the current federation's through KD. In the personalization part, if the common model does not have enough performance on the validation data of the current federation, it refers little to it, while the weight of the KD regularization term is adapted if the common model's performance is acceptable on the current validation data.

Challenge and Motivation: As stated previously, the non-IID nature of local datasets leads to the model drift issue and consequently the knowledge drift problem. The above mentioned works mainly focused to distill knowledge of the teacher(s) into student model(s) based on knowledge-level or model-level information. On the contrary, the proposed KnFu algorithm tries to investigate the effectiveness of local knowledge at the level of local data without having access to local datasets or their information. To this end, we utilize local knowledge of clients to extract specific coarse-grained

information, i.e., an estimation of their class distribution, to determine the semantic neighbors of each client regarding the closeness of the estimated class distributions.

Contributions: Our main contributions can be summarized as follows:

- Proposal of the KnFu algorithm that strategically evaluates and fuses only relevant and beneficial knowledge among clients. This personalized approach ensures that knowledge fusion is tailored to the semantic neighbors of each client, mitigating the risk of model drift caused by non-IID local datasets.
- Introduction of a novel mechanism within the KnFu algorithm to assess the relevance and impact of shared knowledge across clients, ensuring that only effective knowledge contributes to the FL process, thereby preventing the dilution of model performance with non-contributory information.

Comprehensive experiments were performed on MNIST, and CIFAR10 datasets to show the effectiveness of the proposed algorithm in comparison with baseline methods in terms of different metrics. The rest of the paper is organized as follows: Section II formulates the problem and provides required background material to follow developments of the papers. The KnFu algorithm is proposed in Section III, while Section IV presents simulation results and analysis. Finally, Section V concludes the paper.

II. PRELIMINARIES AND PROBLEM STATEMENT

In this paper, we aim to perform a supervised C -class classification task. Let's consider a set of N clients, denoted by $\mathbb{U} = \{u_1, \dots, u_N\}$, in the FL system coordinated by a fusion center. Local datasets of clients, represented by $\mathbb{D}_n = \bigcup_{i=1}^{K_n} \{(x_n^i, y_n^i)\}, n \in \{1, \dots, N\}$, are heterogeneous, where (x_n^i, y_n^i) denotes i^{th} data sample including the input and its ground-truth output, and K_n indicates the size of the local dataset \mathbb{D}_n . Each client u_n aims to train a local Convolutional Neural Network (CNN) model, denoted by $f(\cdot; \Omega_n)$, parameterized by Ω_n . Let $f(x; \Omega) = [f_1(x; \Omega), \dots, f_C(x; \Omega)]$ denote the output of the last layer (softmax) of the CNN model, where $\sum_{j=1}^C f_j(x; \Omega) = 1$, and $f_j(x; \Omega)$ indicates the probability of assigning data sample x to the j^{th} class.

Term $p(\Omega) = [p_1(\Omega), \dots, p_C(\Omega)]$ is defined as the Estimated Probability Distribution (EPD) [15] of assigning new data samples to different classes. After training on local datasets, EPD is calculated over a shared dataset among clients (the transfer set) with K_t data samples, denoted by $\mathcal{D}^t = \bigcup_{i=1}^{K_t} \{(x_t^i, y_t^i)\}$. Generally speaking, the transfer set is a public dataset shared among different clients. In this paper, we have created our transfer sets for various tasks by collecting pieces of clients' local datasets such that the transfer sets contain data samples from all clients and classes. To investigate the local model's bias to different classes, EPD can be computed as the expectation of probability distribution

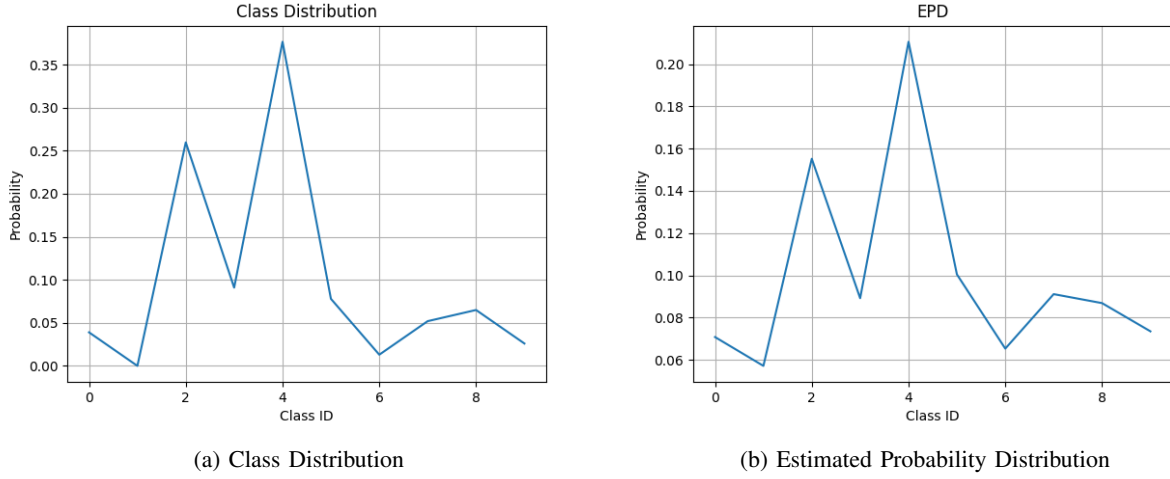


Fig. 1: The EPD as an estimation of the class distribution. The subfigure (a) illustrates an example of local data distribution, e.g. \mathbb{D}_n . The subfigure (b) shows the calculated EPD of the local model that is trained using the local data \mathbb{D}_n . Notably, EPD is calculated with the transfer dataset \mathcal{D}^t after training on the local model \mathbb{D}_n .

of the local models on the data samples of the transfer set, as follows

$$p(\Omega) = \sum_{x \in \mathcal{D}^t} \frac{f(x; \Omega)}{|\mathcal{D}^t|}. \quad (1)$$

KD involves extracting and transferring knowledge from a well-trained teacher model into a student model, emulating the outputs of the teacher model using a transfer set. Specifically, a Kullback-Leibler (KL) divergence function [16] is utilized to minimize the discrepancy between the soft labels of the teacher model and the student model, as follows

$$\Omega_s^* = \underset{\Omega_s}{\operatorname{argmin}} \mathbb{E}_{(x,y) \sim \mathcal{D}^t} \left\{ \mathcal{L}_{CE}(f(x; \Omega_s), y) + \lambda^2 \mathcal{L}_{KL}(f(x; \Omega_s), f(x; \Omega_t)) \right\}, \quad (2)$$

where $f(x; \Omega_s)$ and $f(x; \Omega_t)$ denote the predictions of the student and teacher models for input x , respectively. In addition, \mathcal{L}_{CE} and \mathcal{L}_{KL} are the cross-entropy and KL loss functions, and λ is the so-called temperature hyper-parameter used to soften generated logits. In [6], clients first update their local models using their respective local datasets, then, each client performs predictions on the transfer set to extract a set of soft labels, known as local knowledge. These local soft labels are averaged in the fusion center to fuse the local knowledge of the clients. Finally, the average soft labels, known as collaborative knowledge, are utilized in Eq (2) to distill it into the local models. In [6] and similar papers, the local knowledge of all clients is aggregated and distilled to each local model. The local models are, however, heterogeneous due to the non-IID nature of local datasets. Consequently, their local knowledge is also heterogeneous, i.e., local knowledge of a specific client may not be effective for all other clients in the FL system.

To address the above mentioned issue, in this paper, we aim to answer the followings questions: *Q1: How can we assess the effectiveness of a specific client's local knowledge*

for other clients? Q2: How can we transfer only the effective knowledge and ignore the adverse knowledge of local models? To answer the first question (Q1), we assess the effectiveness of knowledge-sharing and local training options based on two important factors: (i) Data heterogeneity level, and; (ii) Local dataset size. To answer the second question, we propose the innovative Effective Knowledge Fusion (KnFu) algorithm that evaluates knowledge of local models and fuse semantic neighbors' (i.e., clients with similar data distributions) effective knowledge for each client.

III. THE PROPOSED KNFU ALGORITHM

In this section, we present the proposed KnFu algorithm that effectively combines useful knowledge of various clients within an FL framework, ultimately leading to effective personalized local models. Given the inherent diversity of data among clients, it is crucial to acknowledge that not all knowledge is universally beneficial. Hence, there is a necessity for a mechanism to evaluate the relevance and effectiveness of each client's knowledge for others, therefore, preventing the propagation of adverse knowledge. Moreover, crafting an efficient methodology to distill useful knowledge into specific local models is challenging. The proposed KnFu algorithm, consisting of four primary steps, operates over R rounds or until convergence is achieved.

Step 1: Local Training: Initially, individual models undergo training on their local datasets for a set number of local epochs, denoted by E , as follows

$$\Omega_n^* = \underset{\Omega_n}{\operatorname{argmin}} \mathbb{E}_{(x,y) \sim \mathbb{D}_n} \{ \mathcal{L}_{CE}(f(x; \Omega_n), y) \}. \quad (3)$$

Step 2: Local Knowledge Extraction: Following the update of local models, their knowledge is extracted via the transfer set. The knowledge extraction involves obtaining predictions,

Algorithm 1 Pseudocode of the proposed KnFu algorithm

Input: Local datasets $\{\mathbb{D}_n\}_{n=1}^N$ and transfer set \mathcal{D}^t .
for $r = 1, \dots, R$ **do**
 ## Local Training
 for $u_n \in \mathbb{U}$ **do**
 $\Omega_n^* = \operatorname{argmin}_{\Omega_n} \mathbb{E}_{(x,y) \sim \mathbb{D}_n} \{\mathcal{L}_{CE}(f(x; \Omega_n), y)\}$
 end for
 ## Local Knowledge Extraction
 for $u_n \in \mathbb{U}$ **do**
 $F_n = f(x; \Omega_n), \forall x \in \mathcal{D}^t$
 end for

 FUSION CENTER

 ## Effective Knowledge Fusion
 for $u_n \in \mathbb{U}$ **do**
 ## estimation of class distribution (EPD)
 $p(\Omega_n) = \frac{1}{K_t} \sum_{i=1}^{K_t} F_n^{i,j}$
 ## dissimilarities between EPDs
 $d_{n,m} = KL(p(\Omega_n), p(\Omega_m))$
 ## personalized knowledge fusion
 $F_n^{agg} = \sum_{m=1}^N \frac{w_{n,m}}{\sum_m w_{n,m}} \times F_m$
 end for

 ## Local Model Fine-tuning
 for $u_n \in \mathbb{U}$ **do**
 Eq. (10)
 end for
end for
Output: Personalized local models $\{\Omega_n\}_{n=1}^N$

i.e., soft labels, from each local model for the data samples within the transfer set as,

$$F_n = f(x; \Omega_n), \forall x \in \mathcal{D}^t, \quad (4)$$

where F_n constitutes a matrix where each row corresponds to a data sample within the transfer set. The columns of this matrix show the probability distribution for assigning a particular data sample to various classes.

Step 3: Effective Knowledge Fusion: The knowledge extracted from individual clients is transferred to the fusion center for aggregation, leading to the creation of personalized fused knowledge for each client. Initially, we calculate the Estimated Probability Distributions (EPDs) for clients, which indicate the bias of their local models towards various classes. As illustrated in Fig. 1, EPD serves as an estimation of the class distribution within each client's dataset, and is

$$p(\Omega_n) = \frac{1}{K_t} \sum_{i=1}^{K_t} F_n^{i,j}, \forall n \in \{1, \dots, N\}, \quad (5)$$

where $F_n^{i,j}$ indicates the probability of assigning the i^{th} data sample in the transfer set to the j^{th} class using the local model of user u_n . Next, we adjust the weighting of each client's local knowledge for a specific client according to how similar

their EPDs are to that of the client. We measure the distance between two distributions using KL divergence, calculated as

$$d_{n,m} = KL(p(\Omega_n), p(\Omega_m)), \forall n, m \in \{1, \dots, N\}. \quad (6)$$

To determine the importance of the knowledge from client u_m for the client u_n , we evaluate the similarity between their EPDs by taking the inverse of the squared distance calculated in Eq. (6), as follows

$$w_{n,m} = \frac{1}{(d_{n,m})^2}, \forall n, m \in \{1, \dots, N\}. \quad (7)$$

Notably, the weight of the local knowledge is adjusted by a positive constant β , as

$$w_{n,n} = \beta \times \max\{w_{n,m}\}_{m=1}^N. \quad (8)$$

Finally, the personalized aggregated knowledge for client u_n is calculated as

$$F_n^{agg} = \sum_{m=1}^N \frac{w_{n,m}}{\sum_m w_{n,m}} \times F_m, \forall n \in \{1, \dots, N\}. \quad (9)$$

Step 4: Local Model Fine-tuning: In the final step, the personalized fused knowledge is distributed to clients, allowing them to integrate effective knowledge from other clients into their local models. Clients refine their local models by incorporating the aggregated knowledge alongside their local datasets during fine-tuning via the transfer set, as follows

$$\Omega_n^* = \operatorname{argmin}_{\Omega_n} \mathbb{E}_{(x,y) \sim \mathcal{D}^t} \left\{ \mathcal{L}_{CE}(f(x; \Omega_n), y) + \lambda^2 \mathcal{L}_{KL}(f(x; \Omega_n), F_n^{agg}) \right\}, \quad (10)$$

where \mathcal{D}^t is the transfer set, \mathcal{L}_{KL} denotes the KL loss function and λ adjusts the balance between the two terms of the loss function. The first term focuses only on the local dataset, while the second term concentrates on the aggregated knowledge of clients. This completes description of the proposed KnFu, next we present our simulation results and analysis.

IV. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed KnFu scheme through a comprehensive set of experiments analyzing the model's performance under various settings, i.e., different data sizes and heterogeneity levels.

A. Simulation Setup

Datasets: We conduct simulations using two image datasets: MNIST and CIFAR10. We use a Dirichlet distribution to model expected probabilities over a set of categories to account for the varying distribution of local data among clients. Dirichlet distribution is represented as $Dir(\alpha)$, where α adjusts the non-IID-ness degree, i.e., heterogeneity level. As shown in Fig. 2, smaller values of α result in more skewed and, therefore, more non-IID data.

Model Architecture: In our simulations, two distinct CNN architectures are utilized for MNIST

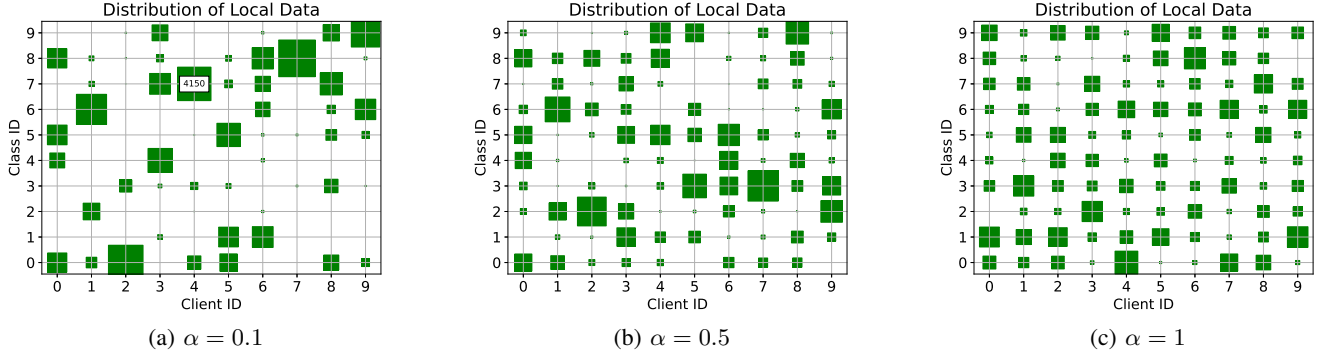


Fig. 2: Illustration of data heterogeneity among 10 clients on the CIFAR-10 dataset.

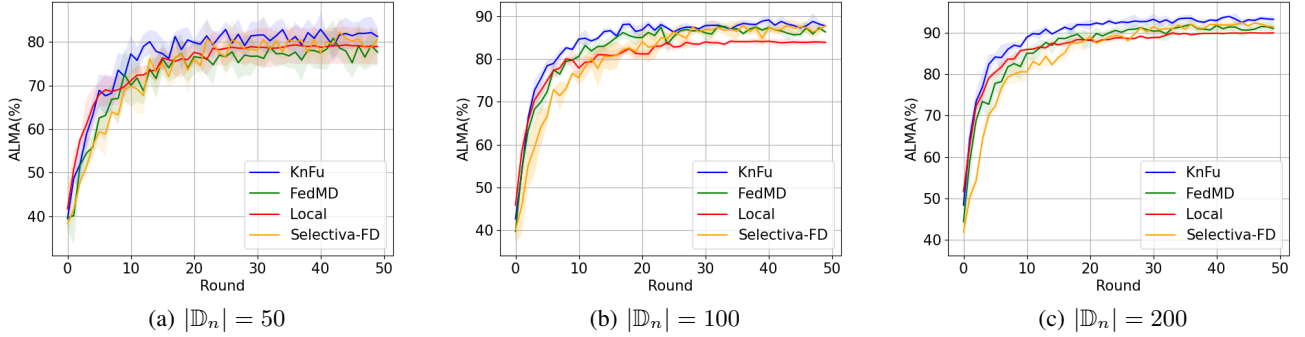


Fig. 3: Learning curves of the ALMA metric corresponding to different methods, for a fixed heterogeneity level, $\alpha = 0.5$, and different local data sizes, $|\mathbb{D}_n| = \{50, 100, 200\}$.

and CIFAR-10 datasets. Specifically, we employ $M1 = [CU_1(32); CU_2(64); FC_1(64); FC_2(32); F_3(10)]$ for the MNIST dataset and $M2 = [CU_1(16); CU_2(16); CU_3(32); CU_4(32); FC_1(128); FC_2(10)]$ for CIFAR-10, where $CU_m(t)$ represents the m^{th} convolutional layer with t channels, and $FC_m(t)$ signifies the m^{th} dense layer with a size of t neurons.

Baselines: The proposed KnFu algorithm is compared with FedMD [6], Selective-FD [10], and local training of the localized models on local datasets (referred to as Local). To ensure fairness across different methods, we maintain the size of the transfer set equal to the local data size. In knowledge fusion-based methods, i.e., KnFu, Selective-FD, and FedMD algorithms, local models are initially updated on their respective local datasets and then fine-tuned on the transfer set using ensemble knowledge. Conversely, in the Local method, both updating and fine-tuning phases occur solely on the local dataset without any knowledge sharing from other clients.

Average Local Model Accuracy (ALMA) serves as the benchmark metric for all methods, reflecting the average test accuracy of all local models on their respective local test datasets. The reported results represent the mean and standard deviation derived from three separate repetitions with distinct random seeds for local model initialization and distinct local data distributions. In each run, the local model initialization and local datasets are the same for all methods.

Hyperparameters: The local epoch is set to $E = 1$ with a batch size of 128, 64, 32, 16, and 8 samples for local data sizes of 1000, 500, 200, 100, and 50, respectively. We consider 20 clients in the FL system. The parameter β in Eq. (8) is set to 10 in all experiments. We execute several experiments for different heterogeneity levels, $\alpha = \{0.1, 0.25, 0.5, 1\}$, and various local data sizes, $|\mathbb{D}_n| = \{50, 100, 200, 500, 1000\}$.

B. Simulation Results and Performance Analysis

In this section, our primary aim is to evaluate the effectiveness of various methods, specifically knowledge-sharing-based algorithms and the local training method, from the perspectives of local data size and heterogeneity level. Fig. 3 demonstrates the impact of different local data sizes on the performance, i.e., ALMA, of various algorithms for a fixed level of heterogeneity. Across all three scenarios, the proposed KnFu algorithm exhibits superior ALMA compared to other baselines. Notably, when the local data size is set to 50, the standard deviation of ALMA for the local training (Local) method varies considerably among different repetitions, although its average ALMA surpasses that of the FedMD algorithm. As the local data size increases, the performance gap between different methods narrows.

Fig. 4 depicts the performance of various methods across different levels of heterogeneity for a fixed local data size. In scenarios of high heterogeneity, i.e., strong non-IID scenarios with $\alpha = 0.1$, knowledge-sharing-based methods do not

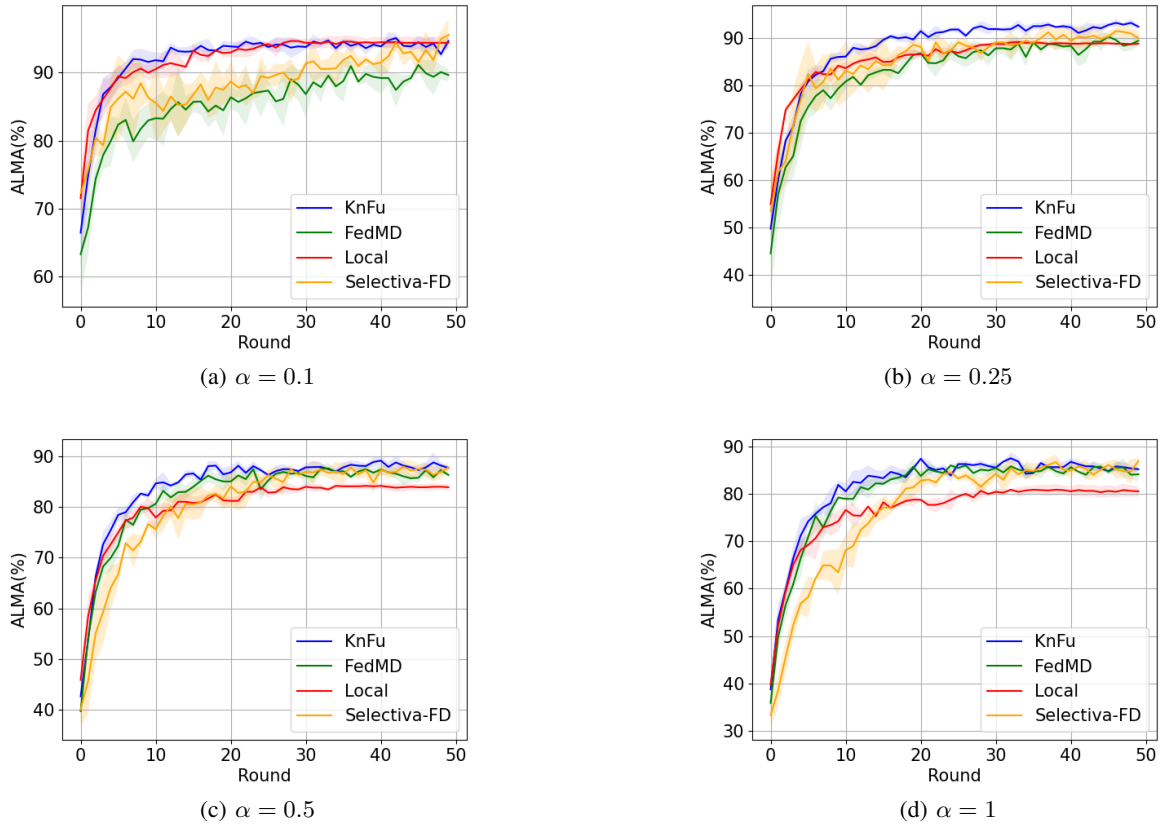


Fig. 4: Learning curves of the ALMA metric associated with different methods, for a fixed local data size, $|\mathbb{D}_n| = 100$, and various heterogeneity levels, $\alpha = \{0.1, 0.25, 0.5, 1\}$.

outperform the local training method. The KnFu algorithm has superior or comparable ALMA compared to the baselines. However, as the non-IID degree of local datasets decreases, the performance of knowledge fusion-based methods converges and surpasses that of the local training method. In summary, in settings with large and highly heterogeneous local data, knowledge fusion algorithms do not offer advantages over the local training method.

Table 1 displays the ALMA (average \pm standard deviation) metric of various methods on the MNIST datasets across different settings, encompassing various local data sizes and heterogeneity levels. Likewise, Table 2 presents the performance of baseline methods on the CIFAR-10 datasets under diverse settings. Unlike the MNIST dataset, the local training method demonstrates superior performance compared to knowledge-sharing-based methods in most scenarios, except for the setting where $\alpha = 1$ and $|\mathbb{D}_n| = 500$. This disparity may stem from the local models' inadequate performance, resulting in the generation of low-quality local knowledge. Similar to the results of the MNIST dataset, it can be observed that in conditions of large and highly heterogeneous local datasets, the local training method is preferable to knowledge fusion-based algorithms. However, the KnFu algorithm outperforms the other knowledge fusion-based methods in most settings.

V. CONCLUSION

In conclusion, FL represents a significant shift from centralized ML approaches, providing a privacy-preserving, decentralized framework for training models across various nodes. Despite its advantages, FL faces challenges such as the requirement for uniform model architectures, and model drift due to non-IID local datasets. While FGD emerged as a solution, leveraging the KD concept to mitigate some of these issues, it fails to fully address model drift, highlighting the need for selective knowledge fusion. The proposed KnFu algorithm offers a novel approach by evaluating and fusing only the relevant knowledge among clients, showcasing superior performance on MNIST and CIFAR10 datasets. This research underlines the potential of personalized knowledge fusion in managing the complexities of FL environments, particularly in the presence of diverse and heterogeneous data. In this work, the KL divergence is utilized to measure the closeness of clients' local knowledge. As a future work, one can investigate the effectiveness of other closeness measure and provide a comparison among them.

REFERENCES

- [1] Khaledyan, D., Marini, T., O'Connell, A., Meng, S., Kan, J., Brennan, G., Zhao, Y., Baran, T. & Parker, K. WATUNet: A Deep Neural Network

TABLE 1: ALMA (%) given different data settings on MNIST dataset.

Het. Level	$ \mathbb{D}_n = 50$				$ \mathbb{D}_n = 100$				$ \mathbb{D}_n = 200$			
	KnFu	FedMD	Local	Selective-FD	KnFu	FedMD	Local	Selective-FD	KnFu	FedMD	Local	Selective-FD
$\alpha = 0.1$	93.5 \pm 0.6	88.7 \pm 0.6	92.4 \pm 2.7	92.2 \pm 1.3	94.1 \pm 0.4	89.3 \pm 0.9	94.4 \pm 0.8	93.1 \pm 1.5	96.2 \pm 0.4	92.2 \pm 0.8	96.3 \pm 0.6	95.9 \pm 0.6
$\alpha = .25$	90.4 \pm 1.3	86.0 \pm 1.1	88.9 \pm 2.1	88.5 \pm 1.7	92.3 \pm 0.6	88.6 \pm 1.2	88.7 \pm 0.2	90.3 \pm 1.1	94.4 \pm 0.7	92.0 \pm 1.3	92.6 \pm 1.1	93.0 \pm 0.9
$\alpha = 0.5$	81.5 \pm 2.7	78.1 \pm 3.6	79.0 \pm 4.0	79.8 \pm 2.6	88.1 \pm 0.6	86.4 \pm 0.2	83.9 \pm 0.6	87.1 \pm 1.0	93.3 \pm 0.5	91.2 \pm 1.1	89.9 \pm 0.0	91.9 \pm 0.9
$\alpha = 1$	78.5 \pm 3.9	76.5 \pm 3.3	75.3 \pm 5.0	78.9 \pm 4.1	85.6 \pm 0.1	84.9 \pm 0.6	80.5 \pm 1.1	85.2 \pm 1.0	92.1 \pm 0.8	92.0 \pm 0.4	87.8 \pm 0.8	92.2 \pm 0.8

TABLE 2: ALMA (%) given different data settings on CIFAR-10 dataset.

Het. Level	$ \mathbb{D}_n = 500$				$ \mathbb{D}_n = 1000$			
	KnFu	FedMD	Local	Selective-FD	KnFu	FedMD	Local	Selective-FD
$\alpha = 0.1$	72.00 \pm 1.57	50.50 \pm 2.51	77.80 \pm 1.78	73.34 \pm 2.47	73.80 \pm 1.23	52.34 \pm 1.18	81.10 \pm 1.81	75.20 \pm 2.32
$\alpha = 0.5$	48.41 \pm 2.30	41.57 \pm 1.94	50.70 \pm 1.25	48.07 \pm 2.13	55.90 \pm 0.98	50.40 \pm 1.68	58.20 \pm 2.21	54.33 \pm 2.45
$\alpha = 1$	45.30 \pm 2.49	42.20 \pm 2.73	43.68 \pm 1.83	44.38 \pm 2.63	48.60 \pm 1.96	46.90 \pm 2.82	50.90 \pm 2.30	48.41 \pm 2.57

for Segmentation of Volumetric Sweep Imaging Ultrasound. *Machine Learning: Science And Technology*. (2023)

- [2] Bayatmakou, F., Mohebi, A. & Ahmadi, A. An interactive query-based approach for summarizing scientific documents. *Information Discovery And Delivery*. **50**, 176-191 (2022)
- [3] Konečný, Jakub and McMahan, H Brendan and Yu, Felix X and Richtárik, Peter and Suresh, Ananda Theertha and Bacon, Dave, “Federated learning: Strategies for improving communication efficiency,” in *arXiv preprint arXiv:1610.05492*, 2016.
- [4] Fu, Lei and Zhang, Huanle and Gao, Ge and Zhang, Mi and Liu, Xin, “Client selection in federated learning: Principles, challenges, and opportunities,” in *IEEE Internet of Things Journal*, 2023.
- [5] Sheikholeslami, S., Rasti-Meymandi, A., Seyed-Mohammadi, S., Abouei, J. & Plataniotis, K. Communication-Efficient Federated Learning for Hybrid VLC/RF Indoor Systems. *IEEE Access*. **10** pp. 126479-126493 (2022)
- [6] Li, Daliang and Wang, Junpu, “Fedmd: Heterogenous federated learning via model distillation,” in *arXiv preprint arXiv:1910.03581*, 2019.
- [7] Atapour, K., Seyedmohammadi, S., Abouei, J., Mohammadi, A. & Plataniotis, K. FedD2S: Personalized Data-Free Federated Knowledge Distillation. *ArXiv Preprint ArXiv:2402.10846*. (2024)
- [8] Chennupati, S., Kamani, M., Cheng, Z. & Chen, L. Adaptive distillation: Aggregating knowledge from multiple paths for efficient distillation. *ArXiv Preprint ArXiv:2110.09674*. (2021)
- [9] Li, Z., Wang, X., Hu, D., Robertson, N., Clifton, D., Meinel, C. & Yang, H. Not All Knowledge Is Created Equal: Mutual Distillation of Confident Knowledge. *ArXiv Preprint ArXiv:2106.01489*. (2021)
- [10] Shao, J., Wu, F. & Zhang, J. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. *Nature Communications*. **15**, 349 (2024)
- [11] Wang, D., Zhang, N., Tao, M. & Chen, X. Knowledge Selection and Local Updating Optimization for Federated Knowledge Distillation With Heterogeneous Models. *IEEE Journal Of Selected Topics In Signal Processing*. **17**, 82-97 (2022)
- [12] Cho, Y., Wang, J., Chirvolu, T. & Joshi, G. Communication-Efficient and Model-Heterogeneous Personalized Federated Learning via Clustered Knowledge Transfer. *IEEE Journal Of Selected Topics In Signal Processing*. **17**, 234-247 (2023)
- [13] Zhang, J., Guo, S., Ma, X., Wang, H., Xu, W. & Wu, F. Parameterized knowledge transfer for personalized federated learning. *Advances In Neural Information Processing Systems*. **34** pp. 10092-10104 (2021)
- [14] Chen, Y., Lu, W., Qin, X., Wang, J. & Xie, X. MetaFed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare. *IEEE Transactions On Neural Networks And Learning Systems*. (2023)
- [15] Abouei, J., Seyedmohammadi, S., Sheikholeslami, S., Mohammadi, A. & Plataniotis, K. MoFLeuR: Motion-based Federated Learning Gesture Recognition. *Authorea Preprints*. (2023)
- [16] Hinton, Geoffrey and Vinyals, Oriol and Dean, Jeff, “Distilling the knowledge in a neural network,” in *arXiv preprint arXiv:1503.02531*, 2015.