

# Goldfish: An Efficient Federated Unlearning Framework

Houzhe Wang<sup>1,2</sup>, Xiaojie Zhu<sup>3(✉)</sup>, Chi Chen<sup>1,2</sup>, Paulo Esteves-Veríssimo<sup>3</sup>

School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

{wanghouzhe, chenchi}@iie.ac.cn

{xiaojie.zhu, paulo.verissimo}@kaust.edu.sa

**Abstract**—With recent legislation on the right to be forgotten, machine unlearning has emerged as a crucial research area. It facilitates the removal of a user’s data from federated trained machine learning models without the necessity for retraining from scratch. However, current machine unlearning algorithms are confronted with challenges of efficiency and validity.

To address the above issues, we propose a new framework, named Goldfish. It comprises four modules: basic model, loss function, optimization, and extension. To address the challenge of low validity in existing machine unlearning algorithms, we propose a novel loss function. It takes into account the loss arising from the discrepancy between predictions and actual labels in the remaining dataset. Simultaneously, it takes into consideration the bias of predicted results on the removed dataset. Moreover, it accounts for the confidence level of predicted results. Additionally, to enhance efficiency, we adopt knowledge distillation technique in the basic model and introduce an optimization module that encompasses the early termination mechanism guided by empirical risk and the data partition mechanism. Furthermore, to bolster the robustness of the aggregated model, we propose an extension module that incorporates a mechanism using adaptive distillation temperature to address the heterogeneity of user local data and a mechanism using adaptive weight to handle the variety in the quality of uploaded models. Finally, we conduct comprehensive experiments to illustrate the effectiveness of proposed approach.

**Index Terms**—federated unlearning, distillation model, efficient retraining

## I. INTRODUCTION

Federated learning is an advanced distributed machine learning paradigm that enables multiple clients to collaboratively train a shared global model without the need to share their local data [1] [2]. This approach effectively addresses a key challenge in traditional machine learning: enabling model training in the absence of centralized storage and processing for datasets. Federated learning effectively addresses this limitation through distributed training, empowering data holders to retain ownership and control of their data while actively contributing to the training of a global model. In this framework, participants are only required to upload the updated

parameters of their local models to the central server. The server then integrates these updates to construct an improved global model. This process serves to safeguard client data privacy.

However, within the federated learning framework, users may request the removal of their contribution from the trained global model. Moreover, recent regulations such as the European Union’s General Data Protection Regulation (GDPR) [3] and the California Consumer Privacy Act (CCPA) [4] empower individuals with the right to demand the deletion of their private data from any part of the system within a reasonable time frame. Furthermore, even if the original data was never shared, the global machine learning model could still glean information about the clients [5] [6]. Predictions made by the global model might potentially leak client information [7] [8]. Therefore, there is a compelling need for a method to eliminate a client’s contribution from the trained global model.

To ensure that a global model forgets the contributions of a specific client, one straightforward approach is to retrain the model from scratch after removing the target user’s data. Alternatively, another method is to remove user information from the parameters of the trained model, while preserving the utility of the model [9].

In the context of managing large datasets and complex models, retraining approaches can be impractical due to high time and energy costs, computational expenses, and scalability challenges. There is a growing interest in developing cost-effective machine unlearning algorithms to mitigate the impact of deleted data from trained models. To measure the effectiveness of an unlearning algorithm in eradicating specified information, Ginart *et al.* [10] introduced a metric similar to  $(\epsilon, \delta)$ -differential privacy (DP). This metric serves as proof of indistinguishability between the output of the unlearning algorithm and the newly retrained output without the deleted records. Building on this concept, several certifications of forgetting have been introduced, aiming to validate the efficacy of various forgetting mechanisms [11]–[15].

However, researchers have demonstrated the weaknesses in current machine unlearning algorithms. Chourasia *et al.* [16] recognizes the interdependence between training data and the model, emphasizing that the data used in the training process

(✉): Corresponding author

Our code is available at <https://github.com/xiao-jian-zi/MU-Goldfish-An-Efficient-Federated-Unlearning-Framework>.

imbues specific patterns within the model. In alignment with Chourasia *et al.*, Gupta *et al.* [17] contend that there are inherent flaws in the deletion certification approach of previous unlearning works, which rely on the indistinguishability between unlearned and retrained models.

Moreover, the existing unlearning algorithms [10]–[12] enhance the efficiency of machine unlearning algorithms by relaxing the requirements for validity. Here, ‘validity’ refers to the effectiveness of the unlearning process in ensuring that the model no longer retains information about the deleted data. Furthermore, most of these algorithms necessitate access to users’ historical model updates or gradient information to enhance efficiency, which is impermissible within the framework of federated learning. Recent studies [18]–[20] have demonstrated that attackers, such as a malicious central server, can exploit clients’ local gradients to mount attacks that reconstruct private training samples of the clients.

In this paper, we aim to overcome the aforementioned shortcomings. First, to address the challenge of low validity in existing machine unlearning algorithms, we propose a novel loss function. It takes into account the loss arising from the discrepancy between predictions and actual labels in the remaining dataset. Simultaneously, it takes into consideration the bias of predicted results on the removed dataset. Moreover, it accounts for the confidence level of predicted results. Additionally, to enhance efficiency, we adopt knowledge distillation technique in basic model and introduce an optimization module that encompasses the early termination mechanism guided by empirical risk and the data partition mechanism. Furthermore, to bolster the robustness of the aggregated model, we propose an extension module that incorporates a mechanism using adaptive distillation temperature to address the heterogeneity of user local data and a mechanism using adaptive weight to handle the variety in the quality of uploaded models.

Overall, the main contributions of this paper are summarized as follows.

- We introduce a novel paradigm for designing a machine unlearning algorithm, named Goldfish. It comprises four modules: basic model, loss function, optimization, and extension.
- We propose a novel design for the loss function that incorporates the discrepancy between predictions and actual labels in the remaining dataset, the bias of predicted results on the removed dataset, and the confidence level of predicted results.
- We propose an optimization module that incorporates the early termination mechanism guided by empirical risk and the data partition mechanism. Additionally, we introduce an extension module that includes a mechanism using adaptive distillation temperature to address the heterogeneity of user local data and a mechanism using adaptive weight to handle the variety in the quality of uploaded models.
- We have established the foundational framework of Goldfish and conducted comprehensive experiments on pub-

licly known datasets to assess its effectiveness. The experimental results demonstrate that the proposed approach can effectively resist backdoor attacks while exhibiting better efficiency and accuracy compared to the state-of-the-art methods.

## II. RELATED WORK

Federated unlearning encompasses two primary approaches: retraining-based and model update adjustment-based methods [21]. Retraining-based methods require substantial retraining of the global model using client data, leading to significant time and resource consumption. In contrast, model update adjustment methods utilize parameter updates from the client to revise the model, eliminating the need for complete retraining. While update adjustment techniques are generally faster and more resource-efficient, they demand the retention of additional information and meticulous calibration and aggregation to ensure efficacy.

### A. Retraining-based approach

Numerous research efforts have been undertaken to expedite the retraining process and minimize time overhead. Liu *et al.* [21] leverages the first-order Taylor expansion approximation technique to customize a rapid retraining algorithm based on diagonal experience FIM. On the contrary, Yuan *et al.* [22] presents a federated forgetting framework. This framework empowers clients to request data deletion, prompting the server to retrain the global model based on these removal requests.

Bourtole *et al.* [9] introduced SISA training as an approach to alleviate computational costs associated with forgetting. This method strategically limits the influence scope of data points during training by employing techniques such as data sharding and slicing.

### B. Model update adjustment-based approach

Much research has been conducted with the aim of enhancing efficiency and effectiveness in the model update adjustment-based approach. Zhang *et al.* [23] eliminate client influence by extracting the weighted sum of gradient residuals from the global model and customizing Gaussian noise. This process is designed to achieve statistical indistinguishability between unlearned and retrained models. Liu *et al.* [24] reconstruct the forgotten model using parameter updates stored on the server, introducing a novel calibration method to adjust client updates. This innovative approach aims to enhance forgetting speed while preserving model performance. In addition, Baumhauer *et al.* [25] and Thudi *et al.* [26] emphasize the pursuit of higher efficiency in machine unlearning by relaxing requirements for both effectiveness and provability. Izzo *et al.* [11], Neel *et al.* [14], and Wu *et al.* [27] explore techniques for the server to effectively approximate gradients during the unlearning process by leveraging historical gradients and model weights. Chourasia *et al.* [28] emphasize the significance of robustness when addressing data deletion scenarios.

The proposed Goldfish framework presents a general approach for realizing federated unlearning, comprising four

TABLE I: Frequently used notations.

$D_c$	the dataset belonging to client $c$
$D_f^c$	the removed dataset of client $c$
$D_r^c$	the remaining dataset of client $c$
$\alpha$	the total number of classes of labels
$\tau$	the total number of shards of user local data
$t$	the number of epochs
$C$	the total number of clients
$\omega$	global model parameter
$\omega_c$	local model parameter

key modules: basic model, loss function, optimization, and extension. The Goldfish framework accommodates the implementation of both retraining-based methods and model update adjustment-based methods. In addition, we introduce a more effective loss function designed to achieve a balance between accuracy and validity in data deletion scenarios. Furthermore, the framework incorporates an adaptive distillation temperature mechanism to address client-side data heterogeneity. Additionally, we propose an adaptive weight mechanism to effectively manage variations in model quality. Finally, to boost the efficiency of the retraining process, we introduce a data sharding mechanism, complemented by an early termination mechanism guided by empirical risk.

### III. PROPOSED FRAMEWORK: GOLDFISH

In this section, we introduce the proposed framework: Goldfish. We first demonstrate the overview of Goldfish and then illustrate its details. Prior to introducing the scheme, we illustrate the commonly used notation in Table I.

#### A. Overview of Goldfish

As shown in Fig. 1, Goldfish specializes in four modules: basic model, loss function, optimization, and extension.

The initial module of the Goldfish, basic model, is crafted for model selection in the context of federated unlearning. The second module is dedicated to the implementation of a personalized loss function. The third module is engineered to execute an optimization algorithm aimed at enhancing efficiency. The final module is tailored to address any additional requirements.

The underlying principle guiding this design is rooted in the characteristics of federated unlearning. In federated learning, as the first step, we need to select the initial model to implement the unlearning process. In particular, the trace generated from the data that is requested to be removed should be eliminated from the global model. To enhance the effectiveness of the selected model, the design of the loss function is pivotal. Loss functions provide a precise metric for assessing a model’s performance by quantifying the disparity between predictions and actual results.

In addition, given the sporadic nature of data removal requests, ensuring the efficient execution of the unlearning process is imperative. To adapt to the evolution of efficient approaches, an optimization module is provided. Lastly, we offer an extension module to achieve compatibility. Additional

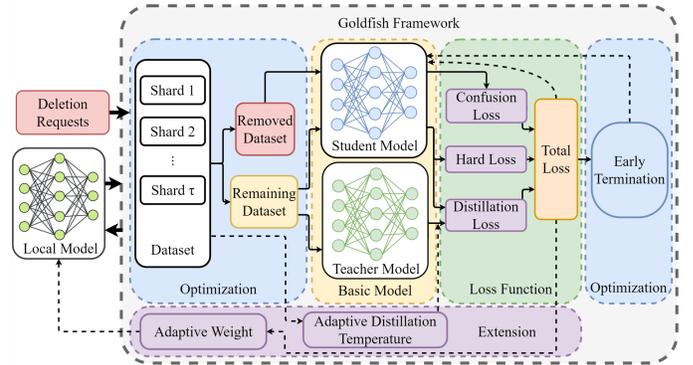


Fig. 1: Goldfish Framework. It consists of four modules: basic model, loss function, optimization, and extension.

requirements can be incorporated into this module. For instance, the approach to address client heterogeneity can be implemented within this module.

#### B. Details of Modules

In this section, we delve into the specifics of each module within Goldfish, and its detailed description of the algorithm is presented in Algorithm 1.

**Basic model.** In Goldfish, the basic model achieves fast retraining through knowledge distillation. In particular, we utilize a teacher model  $M_T$  and a student model  $M_S$  to implement the retraining process on a user’s local dataset  $D_c$ . The local dataset  $D_c$  consists of two parts:  $D_f^c$  and  $D_r^c$ .  $D_f^c$  represents the data that needs to be removed while  $D_r^c$  denotes the remaining data. The student model  $M_S$  can be initialized without any knowledge of  $D_c$ , while the global model serves as the teacher model  $M_T$ . The teacher model  $M_T$  encompasses the knowledge acquired from both datasets  $D_r^c$  and  $D_f^c$ . Our approach is designed to enable the student model  $M_S$  to selectively learn from the teacher model, preserving exclusively the knowledge associated with the remaining dataset  $D_r^c$ , thereby accomplishing the forgetting of  $D_f^c$ . During the training process, we ensure that the knowledge transfer between the teacher model  $M_T$  and the student model  $M_S$  occurs exclusively on dataset  $D_r^c$ , fundamentally preventing the student model from learning the removed data  $D_f^c$ . Moreover, we can employ the removed dataset  $D_f^c$  to evaluate the output of the student model, ensuring that predictions on the removed data are not biased.

**Loss function.** The proposed loss function is an amalgamation of hard loss, confusion loss, and distillation loss.

*Hard loss.* The discrepancy between the model’s prediction and the actual label is named as “hard loss”. The hard loss of the unlearning process on  $D_c$  can be split into two aspects. The first aspect is associated with the remaining dataset  $D_r^c$ . During the training, the label prediction of  $D_r^c$  may deviate from the actual label. The total loss caused by the  $D_r^c$  is denoted as  $L_r$ . Another aspect is related to the removed data  $D_f^c$ . The total loss caused by  $D_f^c$  is denoted as  $L_f$ . Finally, the hard loss is represented by Equation 1. In particular, in multi-class

---

**Algorithm 1: Goldfish Algorithm**

---

**Input :** local training dataset  $D_c$ , initialized model parameter  $\omega_0$ , learning rate  $\mu$ , deletion request

**Output:** unlearned global model

- 1 **Procedure Efficient Federated Unlearning Framework:**
- 2 Reinitialize global model and distribute to all clients  $c$ :  
 $\omega_c^0 \leftarrow \omega_0$ ;
- 3 **for**  $t = 0, 1, \dots, N$  **do**
- 4     **foreach** client  $c$  in parallel **do**
- 5         **if** no deletion request **then**
- 6             LocalTraining( $\omega_c^t, D_c$ );
- 7         **end**
- 8         **else**
- 9             **foreach** unlearned client  $c$  in parallel **do**
- 10                  $D_f^c \leftarrow$  deleted data;
- 11                  $D_r^c \leftarrow D_c / D_f^c$ ;
- 12                 Reinitialize global model and distribute to all clients  $c$ :  $\omega_c^t \leftarrow \omega_0$ ;
- 13                  $\omega_c^{t+1} \leftarrow$ Goldfish( $\omega_c^t, \omega_{t-1}, D_f^c, D_r^c$ );
- 14             **end**
- 15             **foreach** normal client  $c$  in parallel **do**
- 16                  $\omega_c^{t+1} \leftarrow$ Goldfish( $\omega_c^t, \omega_{t-1}, D_f^c, D_r^c$ );
- 17             **end**
- 18         **end**
- 19     **end**
- 20 **Server:**
- 21     Update global model parameters  $\omega_{t+1}$ ;
- 22 **end**
- 23 **return**  $\omega_{t+1}$ ;
- 24 **Procedure** Goldfish( $\omega_c^t, \omega_t, D_f^c, D_r^c$ ):
- 25     Initialize teacher model  $M_T$  parameters:  $\omega_T \leftarrow \omega_t$ ;
- 26     Initialize student model  $M_S$  parameters:  $\omega_S \leftarrow \omega_c^t$ ;
- 27     **foreach** local epoch  $i = 0, 1, 2, \dots, n$  **do**
- 28         **if**  $D_f^c$  is not empty **then**
- 29              $\omega_c^t(i+1) \leftarrow \omega_c^t(i) - \mu \nabla l(M_T, M_S, D_f^c, D_r^c)$ ;
- 30         **end**
- 31         **else**
- 32              $\omega_c^t(i+1) \leftarrow \omega_c^t(i) - \mu \nabla l(M_T, M_S, D_r^c)$ ;
- 33         **end**
- 34     **end**
- 35     **return**  $\omega_c^t(i+1)$ ;
- 36 **Procedure** LocalTraining( $\omega_c^t, D_c$ ):
- 37     **foreach** local epoch  $i = 0, 1, 2, \dots, n$  **do**
- 38          $\omega_c^t(i+1) \leftarrow \omega_c^t(i) - \mu \nabla l(\omega_c^t, D_c)$ ;
- 39     **end**
- 40     **return**  $\omega_c^t(i+1)$ ;

---

classification,  $L_r$  is computed by  $-\sum_{x_i \in D_r^c} y_i \log(M_s(x_i))$ , where  $M_s(x_i)$  represents the prediction made by  $M_s$  on the data  $x_i$  from  $D_r^c$ ,  $y_i$  represents the true label corresponding to  $x_i$ . Similarly,  $L_f$  is computed by  $-\sum_{x_i \in D_f^c} y_i \log(M_s(x_i))$ , where  $M_s(x_i)$  represents the prediction made by  $M_f$  on the data  $x_i$  from  $D_f^c$ . In the paper, we consider the size of  $D_r^c$  is much larger than  $D_f^c$  ( $|D_r^c| \gg |D_f^c|$ ). Without losing generality, the minimization of  $L_h$  promotes  $L_f$  while suppressing  $L_r$ .

$$L_h = L_r - L_f \quad (1)$$

*Confusion loss.* The confusion loss is designed to reduce the bias of the predicted results of  $M_S$  on  $D_f^c$ . We set the Equation 2 to evaluate the bias level of the predicted results,

$$L_c = \frac{1}{|D_f^c|} \sum_{x_j \in D_f^c} \sqrt{D(M_S(x_j))} \quad (2)$$

where  $M_S(x_j)$  represents the predicted vector of the student model for the samples in the removed dataset  $D_f^c$  and  $D(M_S(x_j))$  denotes the variance of the predicted vector. This confidence vector contains the predicted probabilities of a sample belonging to various classes. Variance measures the dispersion between predicted results, and we aim at minimizing it, ensuring that the confidence in predictions made by  $M_S$  for each sample in  $D_f^c$  is close.

*Distillation loss.* Knowledge distillation model utilizes the predicted results generated by the teacher model as labels for the student network. The predicted results of the teacher model are transformed into prediction confidences through the softmax function. The confidence level  $P_{x_i}^T$  of a sample  $x_i$  ( $x_i \in D_r^c$ ) for teacher model  $M_T$  is computed as Equation 3,

$$P_{x_i}^T = \frac{\exp(v_i/T)}{\sum_{j=1}^C \exp(v_j/T)} \quad (3)$$

where  $T$  represents the distillation temperature,  $v_j$  represents the confidence score of the teacher model  $M_T$  predicting  $x_i$  belonging to class  $j$ , and  $v_i$  represents the confidence score of the teacher model  $M_T$  predicting  $x_i$  belonging to the class  $i$  ( $i \in [1, \alpha]$ ) correctly. From Equation 3, we can learn that higher distillation temperatures can make the predicted probabilities generated by the teacher model smoother. Using the same methodology, we define the confidence level  $P_{x_i}^S$  of a sample  $x_i$  for student model  $M_S$  as Equation 4,

$$P_{x_i}^S = \frac{\exp(z_i/T)}{\sum_{j=1}^C \exp(z_j/T)} \quad (4)$$

where  $z_j$  represents the confidence score of the student model  $M_S$  predicting  $x_i$  belonging to class  $j$ , and  $v_i$  represents the confidence score of the student model  $M_S$  predicting  $x_i$  belonging to the class  $i$  ( $i \in [1, \alpha]$ ) correctly.

Finally, the distillation loss is defined based on the confidence level, as formalized by Equation 5. The equation shows that the greater the disparity between the predicted distributions of the teacher model and the student model, the larger the loss.

$$L_d = - \sum_{x_i \in D_r^c} P_{x_i}^T \log P_{x_i}^S \quad (5)$$

*Final expression of loss function.* The final loss function is defined as follows:

$$L = L_h + \mu_c L_c + \mu_d L_d \quad (6)$$

where the weight factors  $\mu_c$  and  $\mu_d$  can be adjusted to balance the importance of different objectives. Assigning a larger weight to the confusion loss can reduce the bias of model  $M_S$  on  $D_f^c$ . The hard loss ensures that the model learns from the remaining dataset  $D_r^c$ , while the distillation loss improves the generalization of the model.

**Optimization.** We introduce two approaches to improve the efficiency of retraining. The first method is implemented through terminating the training process earlier based on convergence speed. The second is to divide the data into many small portions. The retraining process only needs to be conducted on those portions that include deleted data.

*Early termination guided by empirical risk.* During the training process, the loss value output by the loss function is used to measure the convergence of the model. To decide the condition of stopping training, we introduce excess empirical risk to determine whether the loss value of the student model during the training process is close to that of the global model from the previous epoch. The excess empirical risk is defined in Equation 7, where  $\omega^{t-1}$  represents the global model parameters of  $t-1$  epoch of training, and  $w_c^t(i)$  represents the  $i$ th ( $i \in [0, n]$ ) epoch of local training corresponding to the  $t$  epoch of global training, and  $n$  stands for the number of epochs of local training.

$$err(\omega_c^t, \omega^{t-1}) = \left| \frac{1}{n} \sum_{i=0}^n L(\omega_c^t(i)) - L(\omega^{t-1}) \right| \quad (7)$$

During the training phase, when the loss value of the student model decreases to within a certain range of the lowest loss of the pre-trained model, i.e.,  $err(\omega_c^t, \omega^{t-1}) \leq \delta$ , where  $\delta$  is a specified threshold value, the training process terminated.

*Data partition into small portions.* As shown in Fig.2, each local dataset is partitioned into data shards. Each shard has a model and the final output is the aggregation of models from these shards. During the local training process, each data shard is independently trained and retains the model weight  $\omega_{c,i}$  ( $i \in [1, \tau]$ ). Ultimately, all models saved from the data shards are aggregated to construct the user's local model parameter  $\omega_c$ . The above description is formalized in Equation 8, where  $|D_i^c|$  represents the size of the  $i$ th shard and  $|D^c|$  denotes the size of the user local data.

$$\omega_c^t = \sum_{i=1}^{\tau} \frac{|D_i^c|}{|D^c|} \omega_{c,i}^t \quad (8)$$

When a deletion request is made by the user, it is only required to retrain the local model with the remaining dataset by removing the shard that contains the removed data. As a

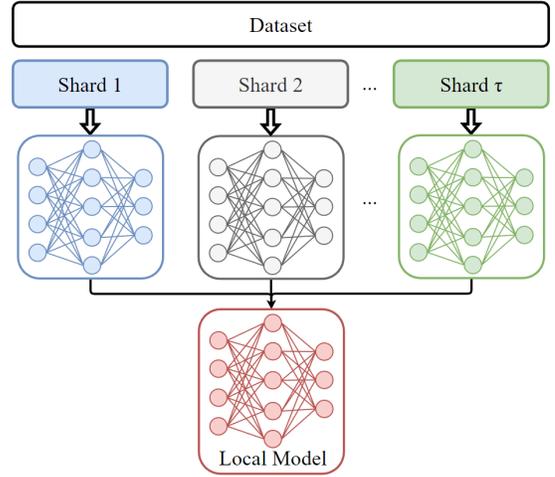


Fig. 2: Data Sharding Diagram. Each dataset is partitioned into data shards. Each shard has a model and the final output is the aggregation of models from these shards.

result, instead of re-initialization of user local model weights, it starts training from the checkpoint, which saves the time of retraining. The checkpoint is computed using the models from data shards that do not contain removed data as Equation 9, where the  $i$ th shard is removed.

$$\omega_c^t = \sum_{j \neq i}^{\tau-1} \frac{|D_j^c|}{|D^c|} \omega_{c,j}^t \quad (9)$$

After retraining, we delete the removed data in data shard  $D_i^c$  and obtain the weights of the shard  $D_i^c$  by subtracting the weights of other shards. As shown in Equation 10, the new model weight of shard  $D_i^c$  is computed by subtracting the model weights of other shards.

$$\omega_{c,i}^t = \frac{|D_i^c|}{|D_i^c|} (\omega_c^{t+1} - \sum_{j \neq i}^{\tau-1} \frac{|D_j^c|}{|D^c|} \omega_{c,j}^t) \quad (10)$$

In case, as shown in Fig.3, if only partial data of a shard is deleted, it is required to retrain the model of the shard. If more than one shard is involved, the training of the multiple shards can be parallelized.

**Extension.** We propose two mechanisms within the extension module. The first mechanism is applied to address the heterogeneity of user local data through adaptive distillation temperature. The second mechanism is employed to tackle the different quality levels of uploaded models via adaptive weights.

*Adaptive distillation temperature proposed for addressing the heterogeneity of user local data.* During the phase of model training, the models trained on clients with larger and more evenly distributed datasets behave better. However, in real-life scenarios, the local datasets vary from user to user. For example, the sizes of local datasets may vary significantly among different users. It is natural to put more emphasis on models that exhibit better performance during model aggregation.

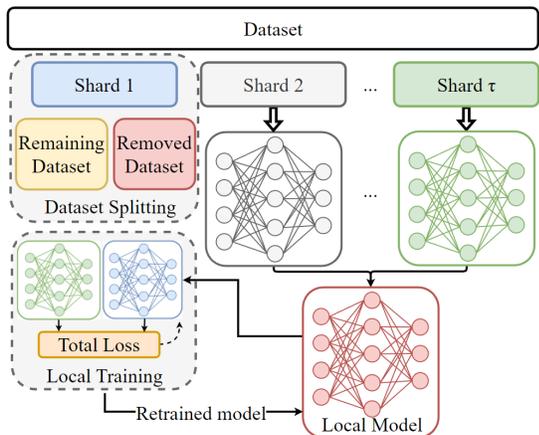


Fig. 3: Retraining under data sharding. In Shard 1, only partial data of the shard is deleted, it is required to retrain the model of the shard before model aggregation.

We formulate the above concept by using Equation 11, where  $T_0$  is the initial distillation temperature, and  $\alpha$  is the adjustment factor. The principle behind the equation is that the amount of information decoupled by the student model from the outputs of the teacher model is directly proportional to the distillation temperature. A higher distillation temperature can smooth the distribution of the teacher model’s predicted results. When  $T \leq 1$ , soft labels degrade into hard labels, and the distillation loss becomes equivalent to the traditional cross-entropy loss. We adjust the value of  $T$  based on the sizes of the removed dataset and the remaining dataset.

$$T = \alpha T_0 \exp\left(\frac{|D_r^c|}{|D_r^c| + |D_f^c|}\right) \quad (11)$$

*Adaptive weight for addressing variety in the quality of uploaded models.* We use the Mean Squared Error (MSE) to measure the performance of different user local models and assign higher weights to models with higher prediction accuracy based on MSE. This approach ensures that the global model retains higher testing performance during model aggregation.

$$\mathcal{W}_c = \exp\left(-\frac{me_c^t - \frac{1}{|C|} \sum_{i=1}^C (me_i^t)}{\frac{1}{|C|} \sum_{i=1}^C (me_i^t)}\right) \quad (12)$$

The detail is shown in Equation 12, where  $\mathcal{W}_c$  denotes the weight allocated to the local model of the client  $c$ ,  $me_c^t$  represents the MSE obtained by the client  $c$  when testing its local model on the test set at the central server,  $\frac{1}{|C|} \sum_{i=1}^C (me_i^t)$  denotes the average of all clients’ MSEs in the  $t$  round of FL. Based on the definition of the client’s weight, in the  $t + 1$ th epoch, the finally aggregated model parameter is defined in Equation 13, where  $\theta$  is the normalization factor, i.e.,  $\theta = \sum_{c=1}^C \mathcal{W}_c$ .

TABLE II: Dataset Description

Dataset	Dimensions	Classes	Training	Test
MNIST	784	10	60000	10000
Fashion-MNIST	784	10	60000	10000
CIFAR-10	3072	10	50000	10000
CIFAR-100	3072	100	50000	10000

$$\omega^{t+1} = \frac{1}{\theta} \sum_{c=1}^C \mathcal{W}_c \omega_c^{t+1} \quad (13)$$

#### IV. PERFORMANCE OF GOLDFISH

In this section, we utilize different datasets and model architectures to evaluate the performance of the proposed forgetting approach. Before delving into the details of the experiment, we provide an overview of the setting below.

##### A. Experimental Setup

**Dataset Description.** In the experiments, we utilized four public known ML datasets: the MNIST [29], Fashion-MNIST(FMNIST) [30], CIFAR-10 [31], and CIFAR-100 [31]. As illustrated in Table II, these datasets encompass diverse attributes, dimensions, and numbers of classes. In the setting of the FL environment, we uniformly assigned the data from the four training datasets to all clients.

**Models.** To enable the variety of models, we adopted four different models. In particular, the model for MNIST and FMNIST is a traditional LeNet-5 model [21] [29] consists of 2 convolutional layers, 2 max pool layers, and 2 fully connected layers in the end for prediction output. The models for CIFAR-10 are a modified LeNet-5 consisting of 2 convolutional layers, 2 max pool layers, and 3 fully connected layers in the end for prediction output and ResNet32, which is a variant of the Residual Network (ResNet) architecture [32]. This model consists of 32 layers, including multiple residual blocks, each designed to learn residual functions with reference to the layer inputs. The model for CIFAR-100 is ResNet56, which is another variant of the Residual Network (ResNet) architecture [32]. This model is designed to handle the increased complexity of the CIFAR-100 dataset. ResNet56 consists of 56 layers, including multiple residual blocks that are instrumental in enabling the network to learn deeper representations without the issues of vanishing gradients that can plague very deep networks. All models are implemented using Pytorch, and experiments are done on a machine with one NVIDIA 1660ti GPU and a machine with one NVIDIA 3060 GPU.

**Hyperparameters.** In the experiment, the total number  $C$  of clients is set to 5, 15, and 25. In addition, the batch size  $B$  is set to 100, and the learning rate  $\eta$  is set to 0.001. For the deletion rate, following [33], we additionally include the values of 2%, 4%, 6%, 8%, 10%, and 12%. The momentum parameter  $\beta$  is set to 0.9 and the total number  $N$  of shards, following [9], is set to 1, 3, 6, 9, 12, 15, and 18.

**Evaluation Metrics.** In the experiments, we evaluate the proposed approach in training speed, accuracy of prediction, unlearning ability, and robustness. In particular, following the

validation approach in [34], we use backdoor attack to evaluate the unlearning ability. Additionally, we employ L2 distance, Jensen-Shannon Divergence (JSD), and T-test to evaluate the model running result.

**Baselines.** To validate the effectiveness of the proposed approach, three baselines are set. The first baseline is an approach to retrain the model from the scratch [23], denoted as  $B_1$ . The second baseline is the rapid retraining method [21], denoted as  $B_2$ . The third baseline is the federated unlearning algorithm [35], denoted as  $B_3$ .

## B. Experiment

The experiments are crafted to assess the effectiveness and forgetfulness of the proposed approaches.

We first execute the retraining algorithms across four distinct datasets. Specifically, we set the batch size  $B$  to 100, and learning rate  $\eta$  to 0.001. In addition, we employed models with accuracies of 94.5%, 81.3%, and 72.6% as teacher models for the datasets MNIST, FMNIST, and CIFAR-100, respectively. Additionally, a modified LeNet-5 model with an accuracy of 72.6% and a ResNet32 model with an accuracy of 92.6% were utilized as teacher models for the CIFAR-10 dataset.. The retraining results for MNIST, FMNIST, CIFAR-10 and CIFAR-100 are illustrated in Fig.4a, Fig.4b, Fig.4c, Fig.4d, and Fig.4e respectively. Upon comparison, it becomes evident that our approach attains the highest accuracy, followed by  $B_2$  in second place, while  $B_1$  exhibits the lowest accuracy.

We examine the success rate of backdoor attacks on models across various deletion rate settings. The objective is to assess whether our proposed approach effectively prevents the retention of pertinent information about backdoor samples in the model after retraining. In our experiments, we compare our proposed method with the original model,  $B_1$  and  $B_3$ . Following the configuration of [36], we set the distillation temperature  $T$  to 3, the weight factor  $\mu_d$  to 1.0, and  $\mu_c$  to 0.25.  $B_2$  speeds up the retraining process by preserving historical gradients and other information, which is the same as  $B_1$ . Both retrain from scratch. Therefore, it is not included here.

The experimental result is shown in Fig.5a, Fig.5b, Fig.5c, Fig.5d, and Fig.5e. From the figures, we can learn that the original model consistently maintains a high backdoor attack success rate, indicating contamination of model parameters and the emergence of specific backdoor prediction patterns. In contrast, our method, as well as  $B_1$  and  $B_3$ , consistently shows low backdoor attack success rates, particularly on MNIST and FMNIST datasets. Our method consistently maintains the lowest backdoor attack success rate across different deletion rates with minimal accuracy error. This underscores the effective prevention of information retention from the removed set during retraining.

In Table III, Table IV, Table V, and Table VI, we present the result of running different models on the test set with various deletion rates. From the tables, it is evident that our proposed method consistently maintains high accuracy across different deletion rates. This indicates the effectiveness of our approach

in ensuring forgetting while preserving predictive accuracy on the test set.

As demonstrated in Table VII, Table VIII, and Table IX, we compare  $B_3$  and our approach with  $B_1$  in terms of JSD and L2 distance on MNIST, FMNIST and CIFAR-10 datasets. Additionally, we conduct a T-test to compare  $B_3$  and our approach with the original model. Jensen-Shannon Divergence (JSD) and L2 Distance are statistical metrics used to quantify the dissimilarity between two probability distributions. JSD measures the average Kullback-Leibler Divergence between the two distributions, whereas L2 Distance measures the mean squared error between them. A small JSD and L2 distance implies a high similarity between two distributions. A T-test is a statistical method used to determine if there is a significant difference between the means of two groups. In a T-test, the p-value (probability value) indicates the probability of obtaining the observed results (or more extreme) if the null hypothesis (*i.e.*, the means of two groups are equal) is true. The smaller the p-value, the stronger the evidence against the null hypothesis, indicating a greater likelihood of a significant difference between the means of the two samples. From the tables, we can learn that both our approach and  $B_3$  have small L2 values. Moreover, our method has a smaller JSD value compared with  $B_3$ , indicating that the predictive results obtained through our method are closer to those obtained through  $B_1$ . In the context of the T-test, our algorithm consistently yields smaller  $p$ -values in most cases. This suggests significant differences between the predictive patterns obtained through our algorithm and those generated by backdoor attacks.

In order to investigate the importance of different components of the loss function, following [36], we conducted an ablation experiment on the loss function. We trained the ResNet32 model on the CIFAR-10 dataset, considering four combinations of loss functions: hard loss only, without distillation loss (*i.e.*, hard loss and confusion loss), without confusion loss (*i.e.*, hard loss and distillation loss), and total loss (*i.e.*, hard loss, distillation loss, and confusion loss) to study the roles of different parts of the loss function. The experimental result is shown in Table X.

After analyzing the experimental data, we observe that the model trained without distillation loss exhibits a low success rate in backdoor attacks while compromising overall accuracy. However, the model trained without confusion loss maintains a high accuracy and success rate in backdoor attacks. Notably, the incorporation of total loss ensures both elevated model accuracy and a reduced success rate in backdoor attacks. This underscores the advantageous role of distillation loss in expediting the training process and highlights the significance of confusion loss in facilitating more effective forgetting of backdoor data during training. Additionally, we note a marginal decrease in accuracy during the initial training stages for the model lacking confusion loss compared to other models. This decline is attributed to the absence of confusion loss, leading the model to adopt incorrect predictive patterns.

Furthermore, we have introduced additional loss functions,

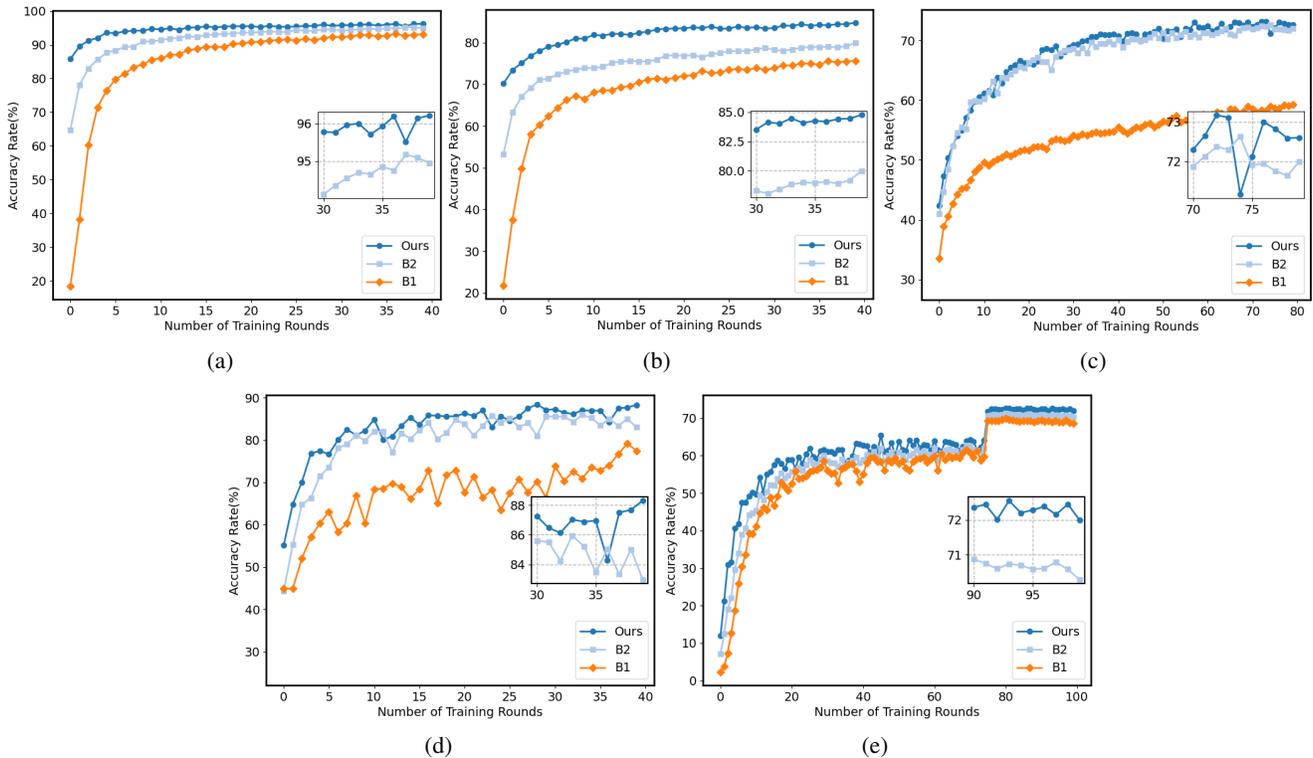


Fig. 4: Accuracy rate of (a) LeNet-5 model trained on the MNIST dataset, (b) LeNet-5 model trained on FMNIST dataset, (c) Modified LeNet-5 model trained on CIFAR-10 dataset, (d) ResNet32 model trained on CIFAR-10 dataset, and (e) ResNet56 model trained on CIFAR-100 dataset.

TABLE III: Accuracy rate and success rate of backdoor attack of models on the MNIST dataset.

Deleted data Rate	origin		Ours		$B1$		$B3$	
	acc	backdoor	acc	backdoor	acc	backdoor	acc	backdoor
2	85.67	89.20	92.67	0.31	95.00	1.90	91.33	1.27
4	86.58	92.79	92.58	0.32	95.50	0.65	90.67	1.29
6	85.94	94.24	95.06	0.10	95.94	0.52	93.61	0.21
8	86.08	96.23	94.25	0.54	95.38	1.08	91.18	0.77
10	85.63	96.98	95.10	0.68	95.63	0.74	95.33	0.74
12	85.33	96.71	94.75	0.41	95.36	0.77	95.56	0.41

namely Focal loss [37] and Negative Log Likelihood (NLL) loss [29], in addition to the traditional cross-entropy loss employed as the hard loss. This sequential inclusion aims to systematically assess the compatibility of our framework with various types of loss functions, as illustrated in the experimental results presented in Table XI. We denote the scenario where the cross-entropy loss function is utilized as the hard loss as ‘Total loss  $\alpha$ ’, while the scenarios employing Focal loss and NLL loss are denoted as ‘Total loss  $\beta$ ’ and ‘Total loss  $\gamma$ ’, respectively. The table demonstrates consistently high accuracy in testing across different loss functions, coupled with a consistently low success rate in backdoor attacks. This underscores the robust compatibility of the proposed framework with diverse loss functions.

To measure the efficacy of data sharding, we seek answers to two pivotal questions:

- Does the model exhibit effective convergence with the application of data sharding?

- Does the utilization of data sharding contribute to an expedited retraining process?

Following the experimental configuration [9], the number of shards is set to 1, 3, 6, 9, 12, 15, and 18. In addition, the MNIST dataset is employed. Initially, each data shard is assigned with a model and each of them is trained over the shard. After that, those models are locally aggregated and the aggregated model becomes the client’s local model. The training result is shown in Fig.6. From the figure, it is evident that the rate of accuracy improvement decelerates with an increasing number of data shards. This phenomenon is attributed to the partiality of data within each shard, resulting in the model being biased towards local data. However, irrespective of the number of shards, the convergence trend of model accuracy remains consistent.

Moreover, for the optimal selection of the number of data shards, it depends on the acceptable accuracy loss and benefits from a reduced number of training rounds in data deletion. In

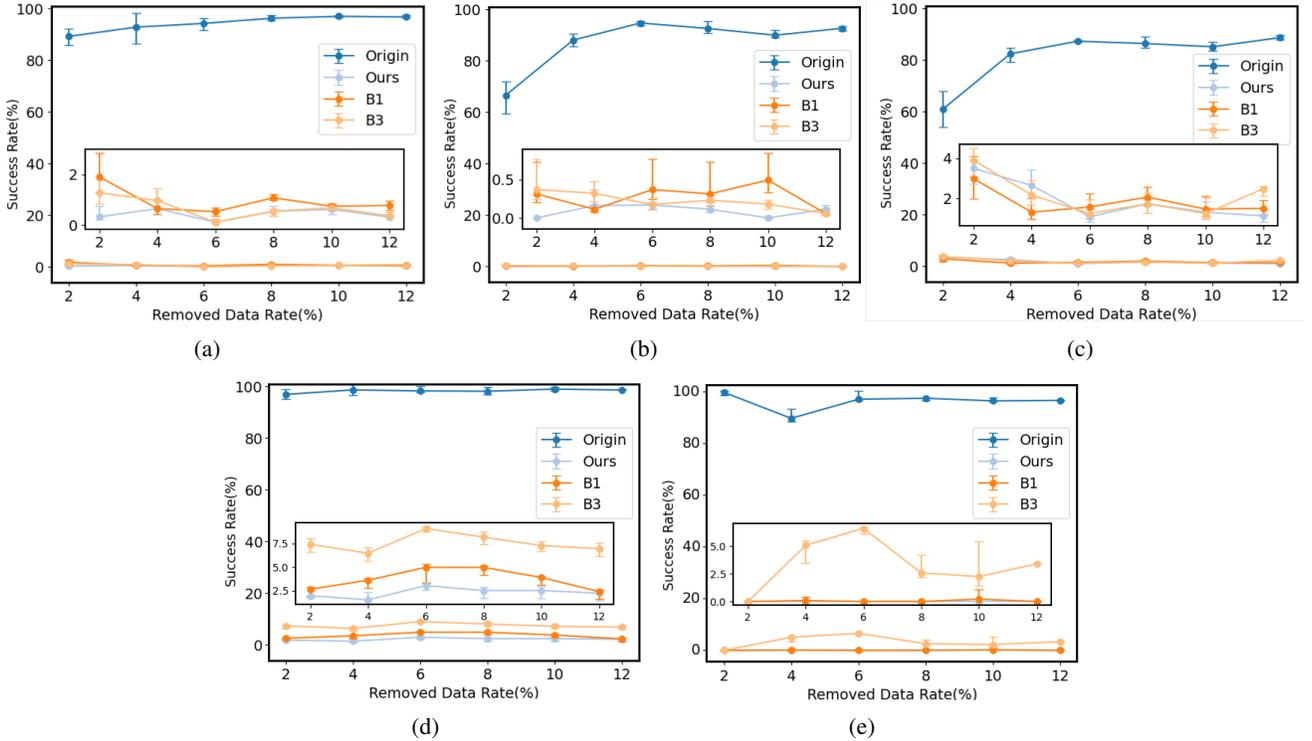


Fig. 5: Success Rate of backdoor attack of models under different removed data rates on the (a) MNIST dataset, (b) FMNIST dataset, (c) CIFAR-10 dataset, (d) CIFAR-10 dataset, and (e) CIFAR-100 dataset.

TABLE IV: Accuracy rate and success rate of backdoor attack of models on the FMNIST dataset.

Deleted data Rate	origin		Ours		B1		B3	
	acc	backdoor	acc	backdoor	acc	backdoor	acc	backdoor
2	75.33	66.56	78.67	0.00	79.50	0.47	73.50	0.31
4	72.50	88.00	76.83	0.16	80.33	0.31	75.17	0.16
6	72.89	94.65	77.17	0.17	80.50	0.00	75.61	0.00
8	71.50	92.52	77.33	0.11	81.33	0.34	74.62	0.17
10	71.90	89.89	78.57	0.00	81.33	0.06	74.13	0.07
12	71.17	92.60	77.58	0.11	78.17	0.06	75.08	0.05

TABLE V: Accuracy rate and success rate of backdoor attack of models on the CIFAR-10 dataset.

Deleted data Rate	origin		Ours		B1		B3	
	acc	backdoor	acc	backdoor	acc	backdoor	acc	backdoor
2	78.54	97.14	87.11	1.90	84.61	2.86	87.53	7.62
4	77.57	99.05	86.31	1.43	85.13	3.81	85.64	6.67
6	75.47	98.10	86.28	3.16	84.69	5.38	87.69	9.18
8	78.65	98.28	86.40	2.71	84.97	5.17	88.72	8.37
10	78.52	99.02	86.21	2.73	84.96	4.10	87.71	7.42
12	78.52	98.54	86.31	2.44	84.97	2.60	86.64	7.15

formal, we could write the above problem as computing the maximum value of  $(rr \cdot c_1 - al \cdot c_2)$ , where  $rr$  is the reduced number of training rounds,  $al$  is the accuracy loss,  $c_1$  is the benefit of one reduced training round, and  $c_2$  is the cost of one percent of accuracy loss.  $rr$  and  $al$  are corresponding to the chosen of the number of data shards while  $c_1$  and  $c_2$  are set based on the user preference.

To investigate the second question, we configured deletion rates at 2%, 6%, and 10% and observed the variations in model accuracy before and after data deletion across different numbers of data shards. The experimental results are depicted

in Fig.7a, Fig.7b, and Fig.7c, with the red dashed line at 3 indicating the moment of the data deletion operation. From Fig.7a, it is clear that when the deletion rate is 2%, the model utilizing data sharding attains higher accuracy after the initial round of training following data deletion, compared to the model without data sharding (i.e., the number of shards is 1). Furthermore, it rapidly approaches the convergence accuracy value before data deletion. This occurs because, in scenarios with a low deletion rate, the deleted data only resides within one of the data shards, thus gaining an advantage over the model without data sharding at the beginning of training.

TABLE VI: Accuracy rate and success rate of backdoor attack of models on the CIFAR-100 dataset.

Deleted data Rate	origin		Ours		B1		B3	
	acc	backdoor	acc	backdoor	acc	backdoor	acc	backdoor
2	57.83	100.00	64.29	0	59.22	0	65.22	0.00
4	55.14	89.47	65.22	0	55.67	0	62.24	5.26
6	51.04	96.67	65.31	0	57.84	0	65.24	6.67
8	53.96	97.62	64.39	0	56.91	0	63.08	2.38
10	50.96	96.30	65.08	0	55.67	0	65.00	1.85
12	51.26	96.61	66.44	0	55.71	0	67.18	3.39

TABLE VII: Evaluation based on JSD, L2, and T-test on the MNIST dataset.

Deleted data Rate	B3			Ours		
	JSD	L2	T-test	JSD	L2	T-test
2	0.69	0.49	0.95	0.56	0.14	0.29
4	0.69	0.27	0.67	0.69	0.15	0.45
6	0.69	0.25	0.74	0.63	0.06	0.33
8	0.63	0.27	0.85	0.63	0.07	0.43
10	0.69	0.06	0.31	0.63	0.07	0.35
12	0.69	0.06	0.28	0.54	0.07	0.35

TABLE VIII: Evaluation based on JSD, L2, and T-test on the FMNIST dataset.

Deleted data Rate	B3			Ours		
	JSD	L2	T-test	JSD	L2	T-test
2	0.63	0.25	0.92	0.63	0.33	0.82
4	0.69	0.27	0.81	0.69	0.32	0.59
6	0.69	0.3	0.71	0.69	0.31	0.67
8	0.69	0.18	0.82	0.69	0.21	0.64
10	0.69	0.29	0.69	0.69	0.22	0.53
12	0.69	0.22	0.69	0.69	0.16	0.58

TABLE IX: Evaluation based on JSD, L2, and T-test on the CIFAR-10 dataset.

Deleted data Rate	B3			Ours		
	JSD	L2	T-test	JSD	L2	T-test
2	0.70	0.35	0.82	0.70	0.34	0.71
4	0.63	0.36	0.85	0.63	0.37	0.73
6	0.63	0.36	0.75	0.63	0.38	0.67
8	0.59	0.23	0.81	0.58	0.26	0.74
10	0.67	0.37	0.79	0.66	0.33	0.63
12	0.63	0.22	0.73	0.63	0.17	0.68

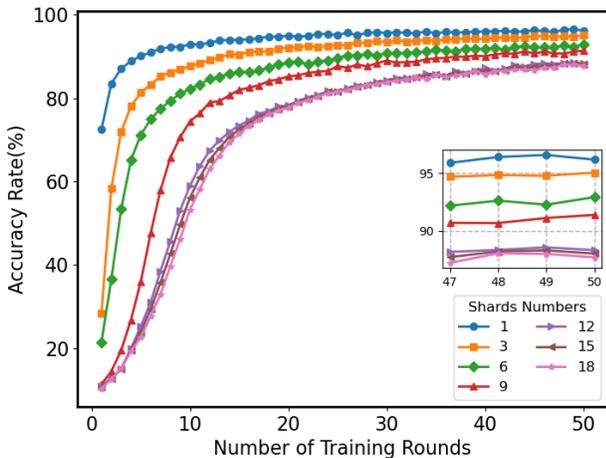


Fig. 6: Accuracy rate of models trained on the MNIST dataset with various numbers of shards.

At a 6% deletion rate, illustrated in Fig.7b, the number of data shards containing deleted data increases. Consequently, models with fewer data shards experience a decrease in accuracy. At a 10% deletion rate, as depicted in Fig.7c, all models utilizing data sharding witness a decline in accuracy. The accuracy with three shards is similar to that with one shard because all shards necessitate retraining when data deletion occurs. However, when the number of shards is 6 and 9, the accuracy remains very high, indicating a trade-off between the number of shards and efficacy. Based on our empirical results, we recommend using a moderate number of shards that allows for efficient retraining without compromising the model’s generalization ability. Specifically, we found that using 6 to 9 shards provided a good balance, as it maintained high accuracy while enabling the model to adapt quickly to data deletions.

Our findings indicate that although the initial advantage in convergence speed is not maintained throughout the entire training process, the data sharding approach shows its resilience against dataset change. Specifically, after experiencing the change in the dataset, the data sharding model can recover faster than a non-sharded model.

We also evaluate the robustness of the proposed model aggregation approach. For benchmarking against the state-of-the-art, we compare our approach with FedAvg [2]. In this evaluation, we aim to answer the following two questions:

- 1) Does our proposed federated model aggregation approach exhibit comparable effectiveness to FedAvg when the client’s local data is independently and identically distributed?
- 2) Does our proposed federated model aggregation approach surpass FedAvg in preserving the performance capability of the model when the client’s local data is highly heterogeneous?

To explore the answers, we set the total number of clients to 5, 15, and 25, and randomly assign the data from MNIST dataset to all users equally.

As shown in Fig.9, the accuracy obtained by our aggregation method and FedAvg aggregation method exhibit virtually identical variations under the condition of uniform data distribution. This indicates comparable applicability for both methods when confronted with uniformly distributed data.

We define heterogeneity as the diversity in datasets across different users. To construct a heterogeneous dataset, data is randomly assigned to each user. To quantitatively evaluate the level of heterogeneity, we calculate the variance of dataset sizes among users and analyze the range of maximum and

TABLE X: Ablation Study of the Importance of Different Components of the Loss Function

Epoch	Metrics	Hard loss only	w/o Distillation loss	w/o Confusion loss	Total loss
10	acc	80.73	80.57	78.91	83.75
	backdoor	5.08	2.93	6.45	3.71
20	acc	84.63	84.19	84.25	85.84
	backdoor	5.08	4.88	7.03	5.08
30	acc	87.03	85.40	86.80	88.22
	backdoor	1.95	2.60	3.13	1.10
40	acc	86.61	86.63	87.71	88.56
	backdoor	5.08	2.93	2.73	2.44

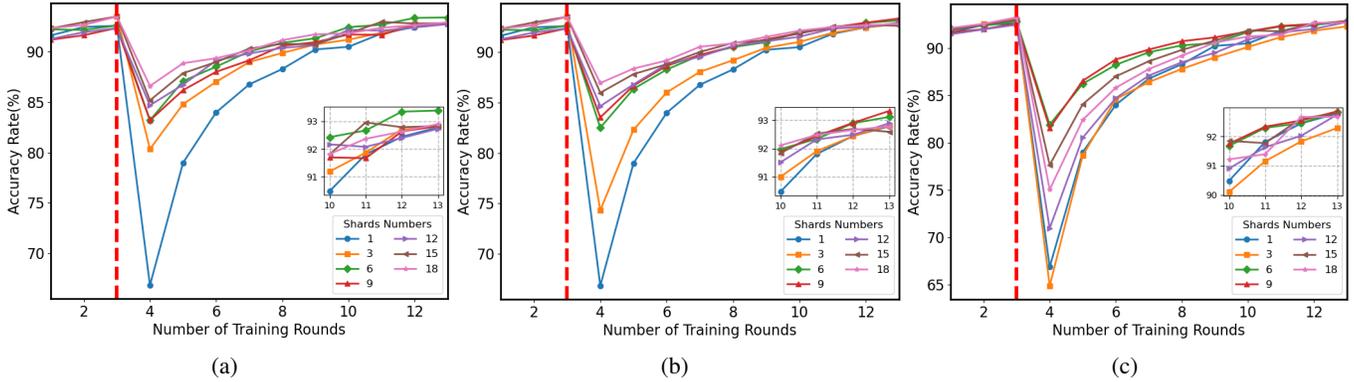


Fig. 7: Model Accuracy on MNIST Dataset with Different Data Shards Numbers at (a) 2%, (b) 6%, and (c) 10% Deletion Rates.

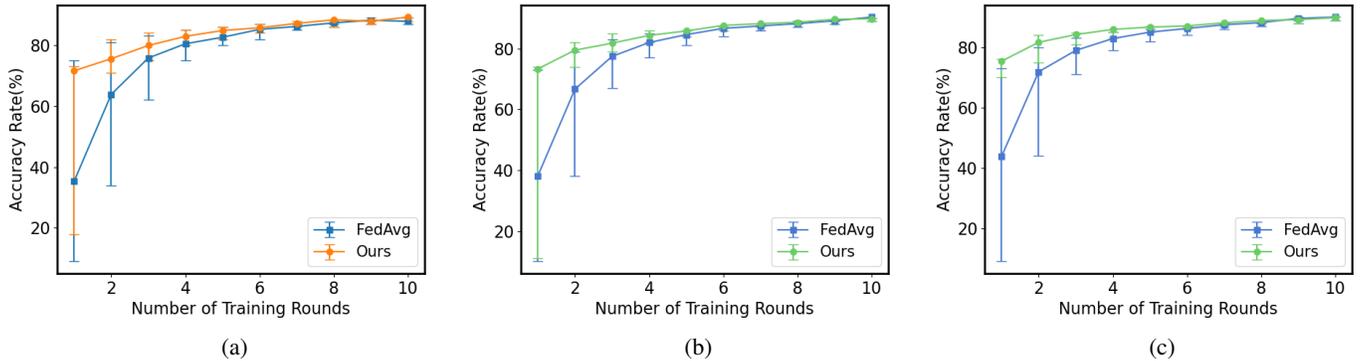


Fig. 8: Accuracy Comparison of FedAvg and Ours on MNIST with Uneven Local Data Distribution for (a) 5, (b) 15, and (c) 25 Clients.

TABLE XI: Compatibility Study of Different Loss Functions.

Epoch	Metrics	Total loss $\alpha$	Total loss $\beta$	Total loss $\gamma$
10	acc	83.75	83.37	83.71
	backdoor	3.71	3.81	2.86
20	acc	85.84	88.03	84.63
	backdoor	5.08	1.9	1.95
30	acc	88.22	87.54	87.71
	backdoor	1.10	1.9	2.86
40	acc	88.56	88.49	87.92
	backdoor	2.44	0.95	2.71

TABLE XII: Representation of Data Heterogeneity.

Clients Numbers	Variance	Min acc	Max acc
5	$2.00 \times 10^6$	9.3	75.6
15	$1.87 \times 10^7$	10.1	73.3
25	$5.20 \times 10^7$	9.8	73.4

minimum accuracy attained by independently training models on local datasets for all users. These metrics are presented in Table XII.

Before performing model aggregation, we assess the ac-

curacy of local models on the test set and compute the accuracy of the global model. The resulting accuracy is plotted with error bars representing precision statistics for all users. This facilitates a comparison of the aggregation capabilities between FedAvg and our proposed method. Figures 8a, 8b, and 8c show wide-ranging error bars of FedAvg at the initial stage due to the heterogeneity of local data among users. This heterogeneity leads to significant variations in the accuracy

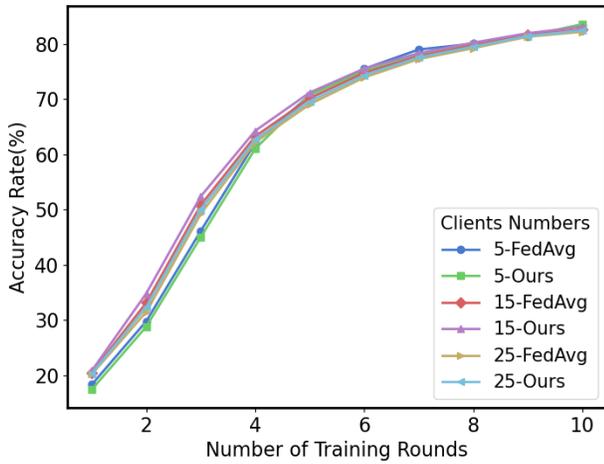


Fig. 9: Accuracy Rate of FedAvg and Ours trained on the MNIST dataset with various numbers of clients. The local datasets for all users are independently and identically distributed.

of local models during the initial training phase. However, our proposed federated model aggregation method effectively addresses the challenge of data heterogeneity in the initial training phase. We assign higher weights to local models demonstrating superior performance on the test set during the aggregation process. This strategic approach allows the global model to retain the parameter distribution of high-performing models, thereby enhancing overall performance and retraining efficiency.

## V. DISCUSSION

In this section, we discuss the challenges and our future research directions of federated unlearning.

The primary challenge lies in striking a balance between the efficiency and validity of federated unlearning. This process aims to expunge specific client data from the global model without compromising model performance. Successfully addressing this challenge demands innovative strategies to streamline the unlearning process while preserving user privacy and maintaining model accuracy. Another challenge is to adapt to a diverse array of heterogeneous clients. In the dynamic landscape of federated unlearning, where clients may join or leave, and data distributions can vary, the federated unlearning scheme must exhibit both flexibility and resilience.

As part of our future research direction, we aim to enhance the accuracy rate of the current data sharding model. Additionally, we plan to explore strategies to effectively address the dynamism and heterogeneity inherent in client environments.

## VI. CONCLUSION

In this paper, we introduce a new paradigm of designing machine unlearning algorithms, named Goldfish. Moreover, we instantiate each module of Goldfish to achieve better efficiency and validity by adopting knowledge distillation technique in basic model, introducing a novel loss function,

proposing an optimization module, and illustrating an extension implementation. Furthermore, we conduct comprehensive experiments to validate the effectiveness of approaches.

## REFERENCES

- [1] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.
- [2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.
- [3] G. D. P. Regulation, "General data protection regulation (gdpr)," *Intersoft Consulting*, Accessed in October, vol. 24, no. 1, 2018.
- [4] S. L. Pardo, "The california consumer privacy act: Towards a european-style privacy regime in the united states," *J. Tech. L. & Pol'y*, vol. 23, p. 68, 2018.
- [5] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning," in *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1–15, 2018.
- [6] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, 2020.
- [7] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.
- [8] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International conference on artificial intelligence and statistics*, pp. 2938–2948, PMLR, 2020.
- [9] L. Bourtole, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, IEEE, 2021.
- [10] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making ai forget you: Data deletion in machine learning," *Advances in neural information processing systems*, vol. 32, 2019.
- [11] Z. Izzo, M. A. Smart, K. Chaudhuri, and J. Zou, "Approximate data deletion from machine learning models," in *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016, PMLR, 2021.
- [12] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," *arXiv preprint arXiv:1911.03030*, 2019.
- [13] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh, "Remember what you want to forget: Algorithms for machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18075–18086, 2021.
- [14] S. Neel, A. Roth, and S. Sharifi-Malvajerdi, "Descent-to-delete: Gradient-based methods for machine unlearning," in *Algorithmic Learning Theory*, pp. 931–962, PMLR, 2021.
- [15] E. Ullah, T. Mai, A. Rao, R. A. Rossi, and R. Arora, "Machine unlearning via algorithmic stability," in *Conference on Learning Theory*, pp. 4126–4142, PMLR, 2021.
- [16] R. Chourasia and N. Shah, "Forget unlearning: Towards true data-deletion in machine learning," in *International Conference on Machine Learning*, pp. 6028–6073, PMLR, 2023.
- [17] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites, "Adaptive machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16319–16330, 2021.
- [18] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pp. 896–911, 2021.
- [19] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.
- [20] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7232–7241, 2021.
- [21] Y. Liu, L. Xu, X. Yuan, C. Wang, and B. Li, "The right to be forgotten in federated learning: An efficient realization with rapid retraining," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1749–1758, IEEE, 2022.

- [22] W. Yuan, H. Yin, F. Wu, S. Zhang, T. He, and H. Wang, "Federated unlearning for on-device recommendation," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 393–401, 2023.
- [23] L. Zhang, T. Zhu, H. Zhang, P. Xiong, and W. Zhou, "Fedrecovery: Differentially private machine unlearning for federated learning frameworks," *IEEE Transactions on Information Forensics and Security*, 2023.
- [24] G. Liu, X. Ma, Y. Yang, C. Wang, and J. Liu, "Federaser: Enabling efficient client-level data removal from federated learning models," in *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pp. 1–10, IEEE, 2021.
- [25] T. Baumhauer, P. Schöttle, and M. Zeppelzauer, "Machine unlearning: Linear filtration for logit-based classifiers," *Machine Learning*, vol. 111, no. 9, pp. 3203–3226, 2022.
- [26] A. Thudi, H. Jia, I. Shumailov, and N. Papernot, "On the necessity of auditable algorithmic definitions for machine unlearning," in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 4007–4022, 2022.
- [27] Y. Wu, E. Dobriban, and S. Davidson, "Deltagrad: Rapid retraining of machine learning models," in *International Conference on Machine Learning*, pp. 10355–10366, PMLR, 2020.
- [28] R. Chourasia and N. Shah, "Forget unlearning: Towards true data-deletion in machine learning," in *International Conference on Machine Learning*, pp. 6028–6073, PMLR, 2023.
- [29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [30] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
- [31] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [33] W. Wang, C. Zhang, Z. Tian, and S. Yu, "Machine unlearning via representation forgetting with parameter self-sharing," *IEEE Transactions on Information Forensics and Security*, 2023.
- [34] C. Wu, S. Zhu, and P. Mitra, "Federated unlearning with knowledge distillation," *arXiv preprint arXiv:2201.09441*, 2022.
- [35] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 7210–7217, 2023.
- [36] X. Zhu, G. Li, and W. Hu, "Heterogeneous federated knowledge graph embedding learning and unlearning," in *Proceedings of the ACM Web Conference 2023*, pp. 2444–2454, 2023.
- [37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Oct 2017.