ConFUSE: Confusion-based Federated Unlearning with Salience Exploration

Syed Irfan Ali Meerza
University of Tennessee
Knoxville, TN, USA
smeerza@vols.utk.edu

Amir Sadovnik
Oak Ridge National Laboratory
Oak Ridge, TN, USA
sadovnika@ornl.gov

Jian Liu
University of Tennessee
Knoxville, TN, USA
jliu@utk.edu

***Abstract*—The increasing scale and complexity of deep neural networks, coupled with heightened privacy concerns, has underscored the importance of developing techniques that align with privacy regulations such as the GDPR and CCPA. These laws mandate the "right to be forgotten", which presents a significant challenge in the context of Federated Learning (FL). FL models trained collaboratively without sharing private data, necessitate efficient unlearning methods that allow for the deletion of specific data without retraining from scratch, which is both computationally and communicatively demanding. This paper introduces a novel framework named ConFUSE, designed to address the multi-faceted challenges of machine unlearning within FL by incorporating neuroscientific principles into a confusion-based technique for memory degradation. This approach enables targeted data erasure at various levels—instance, feature, and client—without the need for knowledge distillation, thus preserving the model's integrity and reducing the computational burden on clients. We evaluate the effectiveness of our method using three benchmark datasets, demonstrating its efficiency and adaptability in FL environments, thereby ensuring compliance with privacy laws and enhancing the model's fairness and reliability.**

***Index Terms*—federated learning, machine unlearning, federated unlearning, model confusion**

I. Introduction

The advent of deep neural networks, with their expansive architectures trained on vast datasets, has significantly advanced machine learning. However, as model sizes balloon and datasets grow, privacy concerns escalate. These sophisticated models are prone to memorizing training data details, posing a stark privacy risk [1]. This memorization runs counter to privacy laws like the GDPR [2] and CCPA [3], which uphold an individual's right to have personal data deleted — a concept known as the *"right to be forgotten"*.

This tension between model performance and privacy rights has catalyzed interest in machine unlearning, which aims to methodically erase the imprint of specific data from models without degrading their utility [4]. Machine unlearning emerges as a formidable task in the realm of Federated Learning (FL), where multiple clients, including many IoT edge devices with limited computational power, collaboratively train a model without sharing their private data [5]. The distributed essence of FL means techniques for machine unlearning developed for centralized architectures cannot be seamlessly adapted. A simplistic approach to federated unlearning would be to retrain the model from the ground up sans the data meant to be omitted. However, this method is impractical in real-world FL scenarios due to the excessive computational and communication demands it places on the system, particularly on the edge and IoT devices. Thus, the quest for efficient unlearning methodologies that align with the distributed, collaborative nature of FL and the dynamic nature of IoT data is not just a technical necessity but a pressing compliance imperative.

Beyond compliance with the *"right to be forgotten"*, unlearning in FL models serves an additional purpose: it allows the system to adapt when training data may be compromised, outdated, or biased over time. Given the decentralized nature of FL and the diverse data contributions from multiple parties, the potential for such data issues is not uncommon. Whether it's due to data poisoning attacks [6], [7], or the simple progression of time rendering certain information less relevant, the capability to selectively unlearn this information is invaluable. It bolsters the security, adaptability, and dependability of FL systems. This proactive unlearning not only aligns with privacy mandates but also underpins the fairness of the FL system. By eliminating data that may skew the model, unlearning ensures that decisions remain just and equitable across all data points.

Recent studies on unlearning within Federated Learning (FL) systems have identified several limitations in the scope and application of current unlearning methods. Existing strategies predominantly focus on *retraining-based unlearning* methods [4], [8], [9], which are often computationally intensive, and *model-revision-based* methods [10]–[14]. Model-revision-based methods mainly address *client-level unlearning* [10], [15], [16], where a client wants to withdraw completely from the federation and seeks to eliminate the impact of all its local data on the global model. However, this focus overlooks the more nuanced need for *instance-level* unlearning, where multiple clients may only want to remove the impact of specific data points from the global model. Moreover, *feature-level* unlearning becomes essential when addressing algorithmic biases that lead to unfair treatment of underprivileged groups. By specifically targeting and unlearning biased features, the fairness and reliability of the FL model can be substantially improved.

Additionally, the majority of contemporary unlearning strategies employ knowledge distillation, where they use the pre-unlearning model as a teacher to transfer retaining knowl-

edge to the post-unlearning model to preserve the model's knowledge [10]–[12], which imposes substantial computational burdens on the clients who may be an IoT device. Many of these algorithms require access to clients' historical model updates or gradient information to improve efficiency — a practice that is often not feasible within FL due to privacy concerns. This restriction further complicates the process of efficiently unlearning targeted data in FL scenarios.

In this paper, we address the practical challenges associated with unlearning in FL and propose a novel framework, namely CONFUSE, for efficient unlearning at various levels — *instance*, *feature*, and *client*. Our approach leverages neuroscience theories on memory forgetting to develop a confusion-based technique that intentionally obscures the model's memory. This is achieved by pairing varied labels with similar feature sets to confuse the model and diminish its recall of the original samples. In addition, our approach avoids the need for knowledge distillation and maintains the model's integrity post-unlearning by using a saliency-guided method. This method decomposes the model into smaller components, allowing for targeted updates that erase specific knowledge without affecting other essential information retained by the model. Our contributions to this work can be summarized as follows:

- We introduce CONFUSE, a versatile framework designed for the nuanced task of unlearning at multiple granularities - individual instances, specific features, or entire client datasets within the FL paradigm.
- Our unlearning process circumvents the traditional reliance on historical updates and gradients, utilizing a confusion-induced method inspired by neuroscientific insights into memory degradation, thus streamlining the unlearning procedure.
- By employing a saliency-guided technique to deconstruct the model into discrete segments, we ensure a precise elimination of targeted knowledge, safeguarding the model's overall acuity and preventing the dilution of unrelated, yet critical, retained information.
- We rigorously assess the efficacy of our method using three benchmark machine learning datasets, demonstrating our approach is efficient and widely applicable within the domain of FL.

II. RELATED WORK

Unlearning in FL is typically implemented through two main strategies: retraining-based and model-revision-based methods. The retraining-based approach necessitates extensive retraining of the global model with client data, while model-revision-based methods adjust the model using client-provided parameter updates, sidestepping the need for retraining.

Retraining-based Unlearning: A considerable body of research has focused on optimizing the retraining process within federated unlearning frameworks. For instance, Liu et al. [8] developed a rapid retraining algorithm that employs first-order Taylor expansion and diagonal experience Fisher Information Matrix (FIM) to reduce time overhead. Yuan et al. [9] introduced a federated forgetting framework that enables clients to request data deletions, prompting the server to retrain the global model accordingly. Bourtoule et al. [4] proposed the Sharded, Isolated, Sliced, and Aggregated (SISA) training method, which minimizes computational costs by limiting the scope of data points' influence through data sharding and slicing techniques.

Model-revision-based Unlearning: On the efficiency front, several researchers have developed methods to enhance the unlearning process in federated settings. Zhang et al. [17] introduced a method to diminish client influence by using a weighted sum of gradient residuals and Gaussian noise, maintaining equivalence between unlearned and retrained models. Liu et al. [13] improved unlearning speed and preserved model accuracy by reconstructing models using server-stored parameter updates and a new calibration method for client updates. Halimi et al. [15] and Wu et al. [16] employed a gradient-based approach to forget data, using the gradient information from the forgetting set. Additional efforts by Baumhauer et al. [18], Thudi et al. [19], Izzo et al. [20] focused on optimizing machine unlearning by developing methods that relax effectiveness standards and improve gradient approximation. Chourasia et al. [14] highlighted the importance of robustness in data deletion, while Wu et al. [10] and Zhu et al. [12] explored knowledge distillation to selectively remove data from models, enhancing the unlearning process in federated learning environments.

III. PRELIMINARIES

A. Federated Learning

Federated Learning (FL) is a decentralized approach to machine learning that enables multiple edge devices, often referred to as clients, to collaboratively train a shared global model without sharing their individual datasets [5]. This methodology helps maintain data privacy while reducing the amount of data transmission required, addressing key concerns in data-sensitive applications. Overall, FL aims to optimize the global objective:

$$\min_{\theta_g} f(\theta_g) = \sum_{k=1}^{K} p_k \mathcal{L}_k(\theta_g),\ \mathcal{L}_k = \mathbb{E}_{(x_i,y_i)\in D^k}[f(\theta_g; x_i), y_i]\} \quad (1)$$

where K is the number of participating clients, each with a participation probability p_k, $\mathcal{L}_k$ is the empirical loss for client k with global model θ_g and $\mathbb{E}$ is the empirical error value.

In a typical FL scenario, each client utilizes its local data to train a global model. Rather than exchanging or centralizing the data, only the model parameters or gradients are shared with a central server. One of the foundational algorithms in this space is FedAvg [21], which aggregates these local models into a global model. The aggregation process involves computing a weighted average of the local models, denoted by $\theta_g^{t+1} = \sum_{k\in K} \frac{n_k}{n} \theta_k^t$, where θ_g^{t+1} represents the parameters of the global model at iteration $t+1$, θ_k^t are the parameters of the local model for client k, n_k is the number of data points at client k, and n is the total number of data points across all clients. FedAvg has demonstrated its ability to effectively

converge even on non-IID (non-Independently and Identically Distributed) data under certain conditions.

B. Federated Unlearning (FU)

Federated Unlearning (FU) has become an essential strategy in FL, facilitating the removal of the influence of specific knowledge (data points, data features, or broader data concepts) from a pre-trained FL model without necessitating complete retraining from scratch. This capability is particularly crucial in federated environments where data privacy and efficiency are paramount. The subset of knowledge designated for removal is known as the *forgetting set*. The primary goal of FU is to update a pre-trained FL model efficiently and effectively so that its performance is comparable to that achieved by full retraining, following the exclusion of the *forgetting set* from the training set.

To illustrate, let $D = \{x_i, y_i\}_{i=1}^{n}$ represent the total training dataset across all the clients comprising n data points, each with inputs x_i where x_i is a collection of features $g_i \in G$ and labels y_i for a supervised learning scenario. Let $D_f \subseteq D$ be the designated *forgetting* set. D_f can contain samples from multiple clients or specific clients depending on the application scenario. In the case of feature-level unlearning, we consider each sample in D_f contains the related feature g_i and the label y_i of the corresponding sample. The complement of D_f, denoted by $D_r = D \setminus D_f$, is known as the *remaining* dataset. Before federated unlearning, the global model, denoted by θ_g, is trained on D using methodologies like empirical risk minimization (ERM) in a federated manner. Retraining is considered the gold standard in unlearning paradigm [13], involving retraining the model parameters θ_g from scratch on D_r. However, model retraining is computationally intensive, presenting a significant challenge in federated settings where resources and bandwidth are often limited. Consequently, the central challenge in FU is to develop an unlearned model θ_u from θ_g using D_f and/or D_r that can accurately and efficiently replace retraining.

IV. Proposed Framework

To address the challenges encountered in FU, we draw insights from cognitive neuroscience theories on memory forgetting. Among these theories, our focus lies on the competitive theory of forgetting [22], which posits that forgetting occurs precisely because memories compete with each other when triggered by the same retrieval cue. This competition can lead to the suppression or inhibition of certain memories, making it difficult to recall them when needed. This theory highlights the dynamic nature of memory retrieval, where multiple memories associated with a retrieval cue compete for activation, and the strongest or most relevant memory tends to dominate the recall process. These competitive dynamics can manifest in two forms: proactive interference, where older memories overshadow new ones upon cue presentation, and retroactive interference, where new memories hinder the recall of older ones.

A. Confusion Loss

Our proposed method, ConFUSE, aligns with the retroactive interference-based competition theory. Specifically, we leverage insights from neural processes on how memories compete within the brain's intricate network of neurons and apply them to the artificial neural networks in the FL. For client k to perform unlearning with its local dataset D^k, D^k is divided into the forgetting set D_f^k and the remaining set D_r^k. We implement the retroactive competing step locally on the client. For each data sample (x_i, y_i) in the forgetting set D_f^k, we create a confusion sample set $(x_i, y_j)|y_j \neq y_i,\ \forall j \in J$ for all the available labels J in the dataset. Combining all the confusion sample sets, we create a confusion set D_c^k. We then compute the confusion loss to optimize the global model:

$$\mathcal{L}_{\text{conf}} = -\log(\sigma(E_{D_f^k}(\theta_g))) - \sum_{D_c^k} \frac{1}{|D_c^k|} \log(\sigma(E_{D_c^k}(\theta_g))) + \sum_{D_c^k} \frac{1}{|D_c^k|} \|E_{D_f^k}(\theta_g) - E_{D_c^k}(\theta_g)\|_2, \tag{2}$$

where $E_{D^k}(\theta_g) = \mathbb{E}_{x_i \sim D^k}[f(\theta_g; x_i), y_i]\}$ and $\sigma(.)$ is the sigmoid function. This loss function encourages the model to "forget" its dependence on the forgetting set D_f^k. It penalizes the model for high confidence in predictions about D_f^k while promoting increased certainty in the confusion set D_c^k, thus inducing a state of confusion regarding D_f^k and reducing predictive accuracy on this subset.

This confusion loss primarily penalizes the model for correctly predicting the samples in the forgetting set, leading to changes in weights associated with these predictions. However, these changes are often localized to specific features and do not necessarily eliminate all useful information the model has learned about the dataset. As a result, the model might still retain subtle knowledge associated with the data in the forgetting set, especially if these patterns are also useful for predicting other data points. This incomplete forgetting process can leave traces in the model's parameters, which can then be exploited. To address this, a regularizer term is introduced to minimize the prediction differences between D_f^k and D_c^k. This function ensures that while the model is forgetting D_f^k, it does not do so by becoming overly confident in its predictions for D_f^k as compared to other subsets. Instead, the model's performance on D_f^k should gently degrade, becoming more in line with its uncertainty about other data points.

B. Saliency-guided Federated Unlearning

After the confusion unlearning, memories of the forgetting set are erased. However, a significant decrease in model performance may happen. The model optimization in the confusion unlearning is limited to specific data samples to forget, which slightly destroys the generalization of the client model. To mitigate this issue, most unlearning-based methods [10], [12] use knowledge distillation to transfer knowledge of the pre-unlearning model to the post-unlearning model. However, this increases the computational overhead for clients who may be an edge device in an IoT network. To mitigate this issue, we use gradient-based weight saliency relying on the

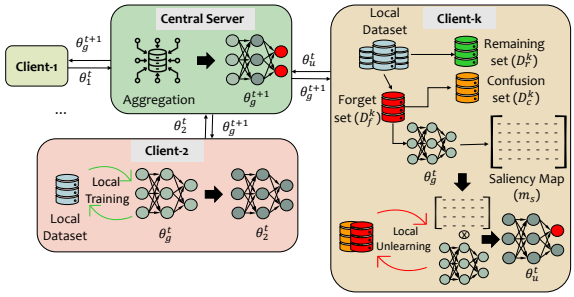


Fig. 1. Illustration of CONFUSE.

fact that contemporary learning models can be decomposed into manageable subparts, each of which can be more easily maintained and updated independently.

Building upon this fact we decompose the pre-unlearning global model weights (θ_g) into two distinct components: the salient model weights earmarked for updating during FU and the intact model weights that remain unchanged. We utilize the gradient of the loss $(l_f^k(\theta_g; D_f^k))$ with respect to the model weights variable (θ_g) under the local forgetting dataset D_f^k of client k. By applying a stochastic thresholding operation, we can then obtain the desired weight saliency mask:

$$m_s = \mathbb{1}\left(\left|\nabla_{\theta_g}(l_f^k(\theta_g; D_f^k))\right| < \frac{1}{1+e^{-\gamma(|\nabla_{\theta_g}|-\tau)}}\right), \tag{3}$$

where $\mathbb{1}(.)$ is an element-wise indicator function that outputs a value of 1 for each model weight if the absolute value of its gradient is below a sigmoid-threshold function and 0 otherwise. The sigmoid function is modulated by a scaling factor γ which adjusts the steepness of the curve, and a threshold τ which shifts the curve along the gradient magnitude axis. The stochastic nature comes from the fact that the sigmoid function introduces a probabilistic "soft" threshold, rather than a hard cutoff. Weights with gradient magnitudes close to τ will have probabilities that could go either way, making the masking process not purely deterministic but rather probabilistic. Leveraging this mask, we articulate the unlearning model for client k as follows:

$$\theta_u^k = m_s \odot (\nabla_{\theta_g} + \theta_g) + (1 - m_s) \odot \theta_g, \tag{4}$$

where $\odot$ is an element-wise product. This equation implies that during the weight update phase, the focus is on updating only the salient weights as identified by the mask, while the remainder of the weights in the model are retained without alteration. This selective focus ensures that the unlearning process is both targeted and efficient, altering only the necessary aspects of the model in response to the removal of the forgetting dataset eliminating the need for knowledge distillation.

C. Design of CONFUSE

Incorporating both the confusion loss and saliency-based unlearning, we introduce CONFUSE that is both effective and computationally efficient. Fig. 1 provides a schematic of our proposed approach. When a client k wishes to unlearn specific knowledge from the global model, it categorizes its dataset into two parts: the forgetting set D_f^k and the remaining set D_r^k. From the forgetting set, the client generates a confusion set D_c^k.

Utilizing the global model, the client first produces a saliency map m_s for the global model weights θ_g using Eq. 3. This saliency map highlights the model weights that are most influenced by the forgetting set. Following this, the client employs both the forgetting set and the confusion set to update the identified salient weights in Eq. 4. This update is governed by the loss function in Eq. 2, focusing on aligning the model's output distributions between the forgetting set and the confusion set. The optimization task for updating the local model is formulated as:

$$\min_{\theta_u^k} \mathcal{L}_{conf} = -\log(\sigma(E_{D_f^k}(\theta_u^k))) - \sum_{D_c^k} \frac{1}{|D_c^k|}\log(\sigma(E_{D_c^k}(\theta_u^k))) + \sum_{D_c^k} \frac{1}{|D_c^k|}\|E_{D_f^k}(\theta_u^k) - E_{D_c^k}(\theta_u^k)\|_2, \tag{5}$$

where $E_{D^k}(\theta_u^k) = \mathbb{E}_{x_i \sim D^k}[\ell(\theta_u^k; x_i), y_i]$ represents the expected loss over the data points in the dataset D^k. This process effectively modifies the local model to "forget" or unlearn the features and knowledge associated with the forgetting set, while the rest of the model remains largely unaffected.

TABLE I
FEDERATED DATASET DESCRIPTION

Dataset	Dimensions	Classes	Clients	Model	FL. Round
MNIST	28 × 28	10	20	LeNet-5	100
CIFAR-10	32 × 32	10	20	ResNet18	100
Adult Income	14 × 1	2	20	MLP	50

V. EVALUATION

Evaluation Scenarios: We evaluate our proposed method through three distinct scenarios to demonstrate its robust capabilities: (i) Neutralizing the influence of backdoor triggers involves client-level unlearning, which is critical for completely removing data from a specific client, especially in situations of data compromise or client withdrawal. (ii) Mitigating the risks associated with membership inference attacks through instance-level unlearning tests the method's precision in selectively forgetting particular data instances. (iii) Eliminating biased triggering features via feature-level unlearning assesses the granularity of our approach in removing specific features. A successful outcome across these scenarios would manifest as diminished model performance on the targeted forgetting data samples, underscoring the versatility and effectiveness of our unlearning approach.

Dataset Description: For our evaluation, we use the following three public datasets that are commonly used in machine learning research: MNIST [23], CIFAR-10 [24], and Adult Income [25]. Details of these datasets, including their attributes, class numbers, and the distribution of data across clients in the federated learning setup, are summarized in Table I. Data from these datasets were uniformly distributed among all clients to simulate a realistic federated learning environment. We set the τ and γ to 0.1 for all the datasets. We initiate the unlearning process after successfully training the model for the mentioned communication rounds in Table I.

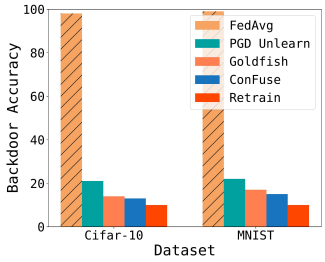


Fig. 2. Backdoor attack success comparison.

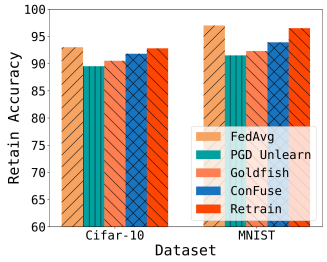
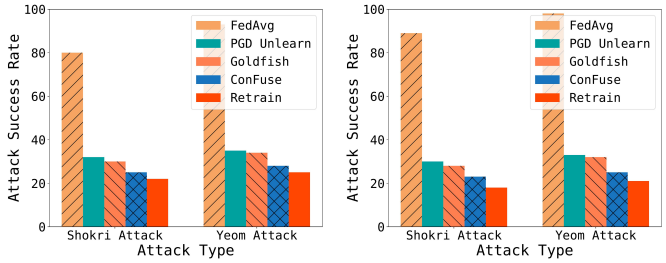


Fig. 3. Non-poisoned set accuracy comparison.



(a) CIFAR-10 (b) MNIST

Fig. 4. Membership inference attacks accuracy (efficacy) for the two attacks.

Evaluation Metrics: Our experimental framework assesses several key performance metrics, including prediction accuracy, attack success rate, true positive rate (TPR), false positive rate (FPR), and training speed. These metrics provide a comprehensive view of the model's performance and the effectiveness of the unlearning process under various conditions.

Baselines: To validate the performance of CONFUSE, we compare it against Four baseline approaches: (i) *FedAvg*, vanilla FL training without any unlearning operations; (ii) *Retrain*, involves retraining the model from scratch, serving as a benchmark for maximum efficacy in removing learned knowledge; (iii) *Goldfish* [11], represents a loss-based unlearning method; and (iv) *PGD Unlearn* [15], employs a Projected Gradient Descent approach tailored for federated unlearning scenarios. These comparisons help to contextualize the performance and advantages of our proposed method within the broader landscape of federated unlearning techniques.

A. Unlearning Under Backdoor Attacks (Client-level)

We use the artificial backdoor triggers [26] as an effective way to evaluate the performance of unlearning methods. Backdoor attacks are uniquely challenging because they do not affect a model's performance on standard inputs but distort predictions when specific, pre-defined triggers are present. This characteristic makes backdoor attacks an ideal test case for evaluating unlearning effectiveness. A successfully unlearned global model should maintain good performance on standard evaluation datasets while significantly reducing the success rate of backdoor attacks. For the experiment, we introduced backdoor triggers into the model by selecting 10% of clients and poisoning their data with a 'pixel pattern' trigger sized 3x3. The selected clients then attempted to unlearn their data samples locally. After three communication rounds of unlearning, we evaluated the global model's performance. The comparative results on backdoor accuracy across various unlearning baselines, including our method CONFUSE, are shown in Fig. 2. Complete retraining sets the benchmark with the lowest backdoor accuracy. The methods *PGD Unlearn* and *Goldfish* lowered backdoor accuracy to 21%, 22% and 14%, 17% for CIFAR-10 and MNIST, respectively. CONFUSE achieved higher reductions, with backdoor accuracies of 13% and 15% on CIFAR-10 and MNIST, closely matching the retraining results.

Beyond backdoor accuracy, maintaining high accuracy on non-poisoned datasets is critical. Fig. 3 details post-unlearning accuracy. While *PGD Unlearn* and *Goldfish* show some accuracy losses — recording 89.5%, 91.5% for CIFAR-10 and 90.5%, 92.3% for MNIST — CONFUSE maintains higher accuracies of 91.8% and 93.9% for the two datasets respectively, demonstrating minimal performance degradation and aligning closely with retraining results.

B. Unlearning Under MIA Attacks (Instance-level)

To further evaluate the efficacy of the method's instance-level unlearning, we conducted a membership inference attack (MIA) test. Unlike the backdoor attack scenario, this setup involved no intentional poisoning of the data samples. We leveraged two prominent membership inference attacks for this assessment: the Shokri attack [27], which constructs shadow models to simulate the target model's behavior, and the Yeom attack [28], which differentiates between members and non-members based on training and test loss values. For this evaluation, we compared the performance of baseline unlearning methods, CONFUSE, a fully retrained model, and the original FedAvg model (the initial global model formed through federated learning before any unlearning efforts). The results of this comparison are illustrated in Fig. 4, which presents the success rates of the MIA across different datasets and attack types before and after unlearning.

The findings show that both baseline unlearning methods significantly reduce the MIA success rate compared to the original FedAvg model. However, these rates are still higher than those observed with the fully retrained model. In contrast, CONFUSE achieved MIA success rates very close to those of the fully retrained model, underscoring the robustness of our approach in enhancing privacy. This demonstrates not only the capability of CONFUSE to effectively mitigate the risks associated with membership inference but also its comparative effectiveness close to that of complete retraining, thereby affirming the high efficacy of our method in data unlearning.

C. Unlearning Under Bias Analysis (Feature-level)

One of the distinctive capabilities of our proposed method, CONFUSE, is its ability to unlearn both specific features and entire instances. While previous evaluations focused on instance-level unlearning, we also assessed feature-level unlearning using the Adult Income dataset. This dataset is particularly suitable for such analysis because it exhibits an inherent gender bias due to a higher proportion of male samples compared to female samples. This imbalance can lead to a model that disproportionately favors one gender. To address this, our unlearning approach specifically targets

TABLE II
PERFORMANCE OF CONFUSE AT BIAS UNLEARNING

Method	TPR		FPR		Accuracy
	Male	Female	Male	Female	
FedAvg	0.88	0.97	0.33	0.46	84.9%
Retrain	0.89	0.87	0.35	0.38	83.7%
CONFUSE	0.88	0.90	0.36	0.41	83.1%

gender features within the dataset. Instead of considering the entire data knowledge as the forgetting set, we selectively target the gender attribute. In practical terms, this means our forgetting set, D_f, consists solely of the gender feature g_i while masking the other features using a binary mask with the labels y_i, making it possible to directly address the bias.

As conventional baseline methods lack support for feature-level unlearning, we compared our approach, CONFUSE, against a model retrained after removing the gender feature from the dataset. We used TPR and FPR to evaluate fairness, revealing that our method effectively reduced gender bias. From Table II we can see that the FedAvg model showed a disparity in sensitivity and specificity between genders. However, after retraining without the gender feature, the model achieved more balanced TPRs of 0.89 for males and 0.87 for females, and FPRs of 0.41 and 0.44, respectively. CONFUSE similarly mitigated gender bias, demonstrating TPRs of 0.88 for males and 0.90 for females, alongside FPRs of 0.40 and 0.45 after unlearning the gender feature. These results affirm the effectiveness of CONFUSE in removing biases and enhancing fairness in machine learning models.

VI. CONCLUSION

In conclusion, our proposed framework, CONFUSE, marks a significant stride in Federated Unlearning (FU) by adeptly addressing at multiple granularities—individual instances, specific features, and entire client datasets. Utilizing a confusion-induced method inspired by neuroscientific insights, our approach moves away from traditional reliance on historical updates and gradients, streamlining the unlearning process while ensuring the precision of memory degradation. The saliency-guided technique we employ allows for the targeted deconstruction and removal of specific knowledge segments, maintaining the integrity and efficacy of the model. Extensive validation on three benchmark machine learning datasets demonstrates that CONFUSE is not only effective but also adaptable across diverse FL scenarios.

REFERENCES

[1] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," *arXiv preprint arXiv:2202.07646*, 2022.

[2] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[3] S. L. Pardau, "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.

[4] L. Bourtoule, 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)*. IEEE, 2021, pp. 141–159.

[5] J. Konečnỳ, 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.

[6] T. D. Nguyen, T. A. Nguyen, A. Tran, K. D. Doan, and K.-S. Wong, "Iba: Towards irreversible backdoor attacks in federated learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] T. Liu, Y. Zhang, Z. Feng, Z. Yang, C. Xu, D. Man, and W. Yang, "Beyond traditional threats: A persistent backdoor attack on federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 359–21 367.

[8] 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*. IEEE, 2022, pp. 1749–1758.

[9] 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*, 2023, pp. 393–401.

[10] C. Wu, S. Zhu, and P. Mitra, "Federated unlearning with knowledge distillation," *arXiv preprint arXiv:2201.09441*, 2022.

[11] H. Wang, X. Zhu, C. Chen, and P. Esteves-Veríssimo, "Goldfish: An efficient federated unlearning framework," *arXiv preprint arXiv:2404.03180*, 2024.

[12] X. Zhu, G. Li, and W. Hu, "Heterogeneous federated knowledge graph embedding learning and unlearning," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2444–2454.

[13] 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)*. IEEE, 2021, pp. 1–10.

[14] R. Chourasia and N. Shah, "Forget unlearning: Towards true data-deletion in machine learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 6028–6073.

[15] A. Halimi, S. Kadhe, A. Rawat, and N. Baracaldo, "Federated unlearning: How to efficiently erase a client in fl?" *arXiv preprint arXiv:2207.05521*, 2022.

[16] L. Wu, S. Guo, J. Wang, Z. Hong, J. Zhang, and Y. Ding, "Federated unlearning: Guarantee the right of clients to forget," *IEEE Network*, vol. 36, no. 5, pp. 129–135, 2022.

[17] 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.

[18] 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.

[19] 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)*, 2022, pp. 4007–4022.

[20] 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*. PMLR, 2021, pp. 2008–2016.

[21] 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*. PMLR, 2017, pp. 1273–1282.

[22] J. T. Wixted, "The role of retroactive interference and consolidation in everyday forgetting," in *Current issues in memory*. Routledge, 2021, pp. 117–143.

[23] 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.

[24] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[25] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[26] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2938–2948.

[27] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[28] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.