FairErase-FCL: Fairness-Aware Federated Continual Unlearning under Non-IID Educational Data

Abstract

Federated unlearning is commonly evaluated by deletion speed, distance to retraining, and average predictive accuracy. These criteria can miss a consequential post-deletion effect: a client withdrawal may alter the distribution of representation across protected groups. Under non-independent and non-identically distributed educational data, the withdrawn institution may contain a disproportionate share of one group. Reversing its update or resetting deletion-sensitive parameters can therefore change group-specific false-negative rates even when average utility remains stable. This paper introduces FairErase-FCL, a fairness-aware sparse repair method for federated continual unlearning. The method identifies coordinates most sensitive to the withdrawn client, builds a group-stratified buffer from data that remain authorized, and optimizes predictive loss together with a smooth equal-opportunity penalty on a restricted repair mask. We evaluate 14 heterogeneous clients, five temporal cohorts, 26 synthetic features, and 15 paired random seeds. The client with the largest protected-group proportion is removed in each run. FairErase-FCL achieves 78.96% ± 1.85% balanced accuracy, 0.42 percentage points below full retraining. Its equal-opportunity gap is 6.65 percentage points, significantly lower than the 8.05 points produced by full retraining ($t(14) = -4.55$, $p < 0.001$), at an estimated 24% of retraining computation. Worst-group accuracy does not improve, revealing a genuine conflict between fairness objectives. The study shows that post-deletion release criteria should report average utility, deletion evidence, and several group metrics rather than accepting a single aggregate accuracy.

Keywords: federated unlearning; continual learning; algorithmic fairness; non-IID educational data; equal opportunity; post-deletion governance

Research highlights

- Defines deletion-induced fairness shift as a first-class risk in federated educational models.
- Repairs only deletion-sensitive and fairness-sensitive coordinates, limiting computation and unrestricted re-learning.
- Uses a group-stratified buffer composed exclusively of data that remain authorized after the withdrawal.
- Reports balanced accuracy, worst-group accuracy, equal-opportunity gap, retraining distance, and cost to avoid a single-metric fairness narrative.

1. Introduction

Educational prediction systems are used for early warning, resource recommendation, course placement, and learning-outcome assessment. Their data are inherently distributed and heterogeneous. Schools differ in student populations, curriculum, devices, assessment practices, participation rates, and feature definitions. Federated learning allows institutions to train a shared model without centralizing raw records [1], but the standard weighted-average objective can prioritize clients with more samples, less noise, or more frequent participation.

Research on fair federated learning has shown that aggregation is not distributionally neutral. q-Fair Federated Learning reweights high-loss clients to reduce variation in client utility [2]. More recent methods directly track group fairness across distributed data [3,4]. These approaches generally assume a training population that remains available. Machine unlearning changes that population by design. When an institution or authorization batch is removed, the retained model is evaluated on a different representation structure.

Consider a client that contains a large proportion of students from a protected or underrepresented group. Its withdrawal may be legally required and correctly executed, yet the remaining global model may become less sensitive to students from that group. A deletion procedure that preserves overall accuracy can still increase the difference in true-positive rates. Conversely, retaining withdrawn records to preserve representation would undermine the deletion objective. Federated unlearning therefore creates a three-way tension among removal, utility, and group fairness.

The supplied course architecture uses task-specific masks to make knowledge addressable. FairErase-FCL retains that sparse-boundary principle but shifts the research question from control-plane throughput to the distributional consequences of deletion: **Can a model use a small, currently authorized, group-stratified buffer to reduce a deletion-induced equal-opportunity gap while keeping repair cost well below full retraining?**

The paper contributes: (1) a formulation of post-deletion fairness as part of the unlearning state transition; (2) a sparse repair method that combines deletion-sensitive and fairness-sensitive coordinates; (3) a release gate that reports plural fairness and utility metrics; and (4) a reproducible experiment that identifies both an improvement in equal opportunity and a cost in worst-group utility.

2. Related work

2.1 Federated heterogeneity and client utility

FedAvg trains local models and aggregates their updates at a coordinating server [1]. Under non-IID data, local optimization can drift away from the global descent direction. SCAFFOLD corrects this effect with control variates [5]. q-FFL changes the objective itself, assigning greater weight to clients with larger losses and reducing the variance of client accuracy [2]. These methods demonstrate that an aggregation rule determines whose errors receive priority. Neither standard FedAvg nor client-level utility fairness directly controls disparities between protected groups distributed across clients.

2.2 Group fairness in federated learning

Rychener et al. propose function tracking for global group-fairness regularization under distributed data [3]. Pentyala et al. examine privacy-preserving group fairness in cross-device federated learning [4]. Both address a key tension: estimating a group-level constraint normally requires sensitive attributes across the federation, whereas federated learning aims to keep such data local. Secure aggregation, local sufficient statistics, and privacy-preserving function tracking can reduce exposure, but the legitimacy and governance of sensitive-attribute use remain application-specific.

2.3 Choosing a fairness criterion

Equality of opportunity requires equal true-positive rates across protected groups [6]. In an educational early-warning setting, the positive class may represent a student who genuinely needs support. A true-positive-rate gap then captures differences in timely identification. The criterion does not constrain false-positive rates, calibration, individual fairness, or treatment quality. It can also conflict with overall accuracy or worst-group accuracy. This paper predeclares equal opportunity as the primary disparity metric and reports balanced accuracy and worst-group accuracy as secondary outcomes.

2.4 Federated unlearning

FedEraser reconstructs a global model by calibrating retained historical client updates [7]. SIFU provides a sequential informed approach with formal guarantees for federated client deletion [8]. Empirical work shows that federated unlearning methods vary substantially in both efficiency and effectiveness [9]. Sparse task isolation, including Privacy-Aware Lifelong Learning and hard task masks, makes a task-level contribution easier to locate [10,11]. Existing evaluations nevertheless emphasize average performance and deletion efficacy. They rarely ask whether the deletion event changes group-specific errors.

3. Problem formulation

Let client c hold samples (x,y,a), where $a0,1$ is a sensitive attribute used only for authorized fairness estimation. Let θ be the trained global model. When client c_r withdraws, the ideal retraining reference is

$$\theta^{-r} = A\left(\bigcup_{c \in C \setminus \{c_r\}} D_c\right).$$

An unlearning algorithm returns $tilde\theta=U(\theta,D\_c\_r)$. Balanced accuracy is

$$\mathrm{BAcc} = \tfrac{1}{2}(\mathrm{TPR} + \mathrm{TNR}).$$

The equal-opportunity gap is

$$\Delta_{EO} = \left| P(\hat{Y} = 1 \mid Y = 1, A = 1) - P(\hat{Y} = 1 \mid Y = 1, A = 0) \right|.$$

Post-deletion repair should not merely minimize the distance to $\theta \wedge -r$ because full retraining can reproduce an undesirable disparity. FairErase-FCL instead solves a constrained behavioral objective:

$$\min_{\tilde{\theta}} \mathcal{L}_{ret}(\tilde{\theta}) + \lambda_f \widehat{\Delta}_{EO}^{\prime 2} + \lambda_s \|\tilde{\theta} - \theta_{reset}\|_M^2.$$

Here, L_ret is computed only from retained authorized data; M is the set of coordinates that the repair procedure may update. The objective explicitly chooses equal opportunity. It does not imply demographic parity, equalized odds, calibration, or universal fairness.

4. The FairErase-FCL method

4.1 Deletion-sensitive mask

The system first estimates the withdrawn client's gradient contribution g_r at the current model. The coordinates with the largest absolute contribution define the deletion-sensitive mask:

$$M_r = \mathrm{TopK}(|g_r|, k).$$

A sparse reverse correction on M_r produces θ_reset. This is an influence approximation rather than a formal deletion guarantee. A deployment requiring provable client-level unlearning should combine the repair stage with a certified method such as SIFU or a complete retraining path [8].

4.2 Authorized group-stratified buffer

The repair buffer $B=B\_0\ B\_1$ is sampled from retained clients, with an equal maximum count for each group. It must satisfy four governance constraints: current authorization, documented repair purpose, exclusion of the withdrawn client, and deletion after the release gate. Stratification does not claim that the production population is balanced. It ensures that a fairness gradient can be estimated from a small buffer without one group dominating the calculation.

4.3 Fairness-sensitive coordinates

For positive examples in the repair buffer, the method computes the difference in mean predicted probabilities:

$$\widehat{\Delta}_{EO}^{soft} = \left| \frac{1}{|B_1^+|} \sum_{i \in B_1^+} p_i - \frac{1}{|B_0^+|} \sum_{i \in B_0^+} p_i \right|.$$

Coordinates with the largest gradient contribution to this disparity form *M_f*. The final writable set is *M=M_r M_f*. All other parameters remain frozen. The model minimizes weighted cross-entropy plus the smooth equal-opportunity penalty within this sparse set. Restricting the writable set reduces computation and limits the chance that broad fine-tuning reconstructs the removed client contribution.

4.4 Release gate

A production release should require all of the following:

1. Declared deletion evidence or retraining proximity is satisfied.
2. Balanced accuracy remains above an approved utility floor.
3. Equal-opportunity gap, worst-group accuracy, false-positive disparity, and calibration are reported together.
4. Each group has sufficient positive examples for a stable estimate.
5. A human reviewer confirms the legal basis for sensitive-attribute processing and the availability of an appeal route.

If the sample count for a group is too small, the system must not publish a seemingly precise disparity value.

5. Experimental design

5.1 Synthetic federated data

The experiment contains 14 clients and five temporal cohorts. For every client-cohort pair, the generator creates 240 training samples and 150 test samples with 26 features. Each client's group proportion is drawn from Beta(1.3, 1.3) and clipped to the interval [0.06, 0.94], producing substantial non-IID composition. Group membership shifts the means of the first two features and the label intercept. Temporal cohorts add gradual concept drift. Labels are sampled from a noisy logistic model.

5.2 Withdrawal scenario and comparators

In every run, the client with the highest proportion of group 1 is removed. This stress test represents a withdrawal that causes concentrated representation loss. Five methods are compared:

- **No unlearning:** the original model is evaluated after the request without modification.
- **Influence step:** a single reverse-gradient approximation is applied.
- **Sparse reset:** only the top deletion-sensitive coordinates are corrected.
- **FairErase-FCL:** the sparse reset is followed by fairness-aware repair.
- **Full retraining:** the model is rebuilt on all retained clients.

FairErase-FCL samples at most 550 records from each retained group, updates no more than 20 combined sensitive coordinates, and performs 220 repair steps. The experiment uses 15 paired random seeds from 6200 to 6214.

5.3 Outcomes and statistics

The primary outcome is the equal-opportunity gap. Secondary outcomes are balanced accuracy, worst-group accuracy, normalized parameter gap to full retraining, and relative computation. Paired *t* tests compare FairErase-FCL with the baselines. Because the study is a method-screening simulation rather than a confirmatory educational trial, *p* values are descriptive and should be read alongside effect sizes and uncertainty.

5.4 Reproducibility

The generator and algorithms are implemented in analysis/simulate_studies.py. Raw and aggregated outputs are stored in analysis/results/study2_raw.csv and analysis/results/study2_summary.csv.

6. Results

Method	Balanced accuracy	Standard deviation	Worst-group accuracy	Equal-opportunity gap	Parameter gap	Relative compute
No unlearning	79.41%	1.85%	79.33%	8.00 pp	0.013	0.0%
Influence step	79.37%	1.83%	79.28%	8.11 pp	**0.009**	1.2%
Sparse reset	79.38%	1.83%	79.28%	7.96 pp	0.036	1.8%
FairErase-FCL	78.96%	1.85%	78.89%	**6.65 pp**	0.149	24.0%
Full retraining	**79.38%**	1.84%	**79.30%**	8.05 pp	0.000	100%

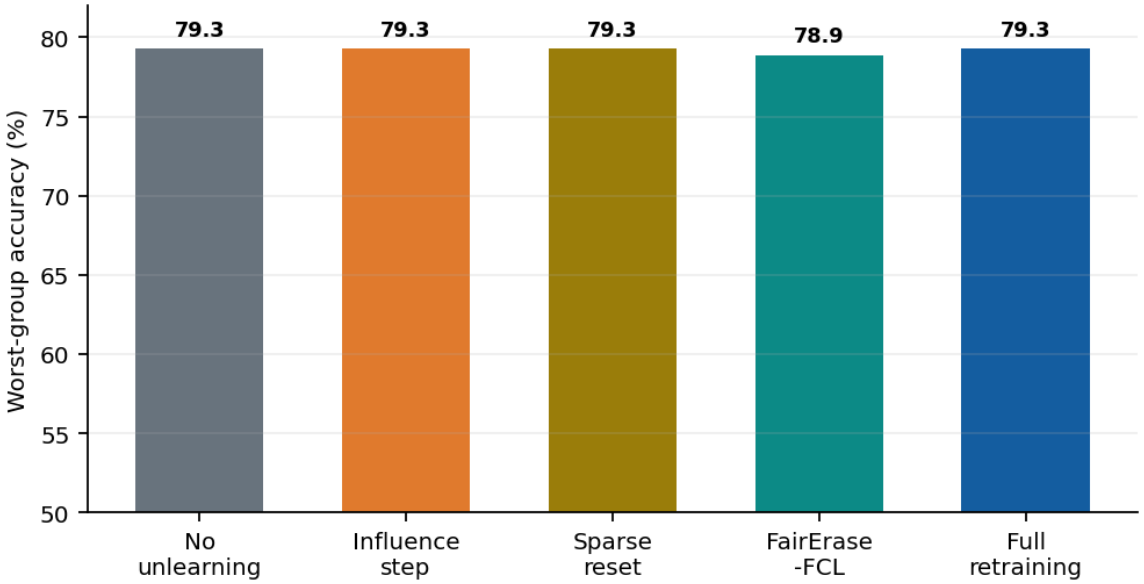

Figure 1. Worst-group accuracy after withdrawal of the client with the most concentrated group representation. FairErase-FCL does not outperform retraining on this metric.

FairErase-FCL reduces the equal-opportunity gap by 1.40 percentage points relative to full retraining, a 17.4% relative reduction. The paired result is $t(14) = -4.55$, $p = 0.00046$. Relative to sparse reset, the reduction is 1.31 points, with $t(14) = -4.18$ and $p = 0.00092$. Balanced accuracy is 0.42 points lower than full retraining, and worst-group accuracy is 0.41 points lower.

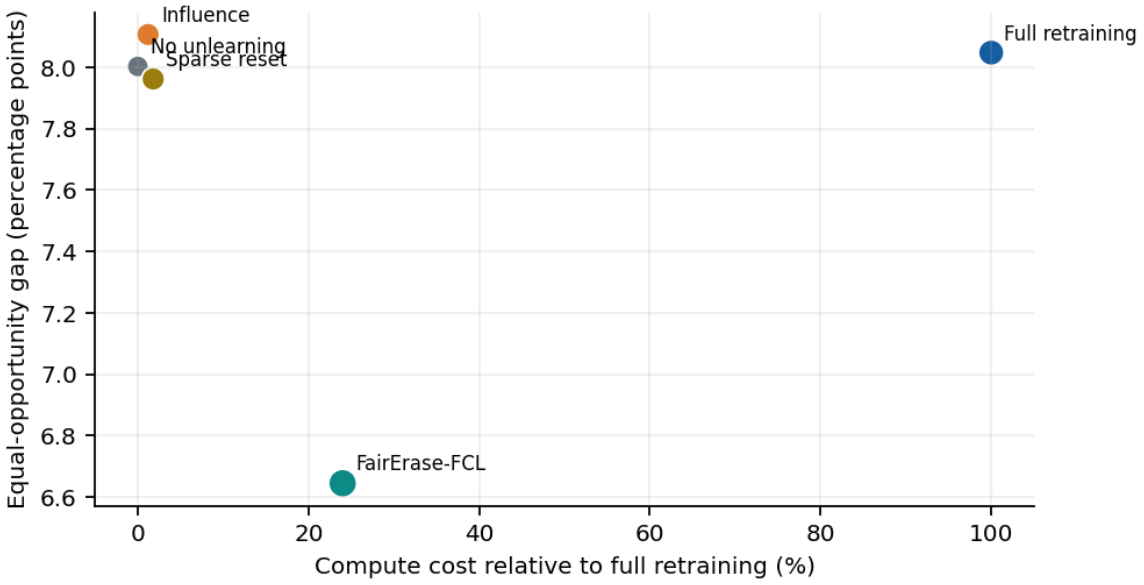

Figure 2. Equal-opportunity gap versus estimated computation. The lower-left region is desirable, but no method dominates on every outcome.

The parameter results reinforce the need for behavioral evaluation. The influence step is closest to full retraining in Euclidean parameter distance but does not improve the fairness metric. FairErase-FCL is farther from retraining yet produces a smaller equal-opportunity gap. Parameter similarity therefore cannot replace direct group-level auditing.

7. Discussion

7.1 Deletion as a governed distribution shift

Conventional drift arises from time, behavior, devices, or curriculum changes. Deletion-induced drift is triggered by a change in authorization. It is foreseeable, attributable, and governed by policy. Treating the withdrawal as an ordinary gradient update ignores the potential loss of group representation. Treating it as a state transition allows the institution to assess affected clients, tasks, and groups before publication.

7.2 Fairness objectives are plural

Equal opportunity is relevant when the main harm is failing to identify students who genuinely need support. In other applications, false positives may produce stigma, unnecessary intervention, or resource diversion. Equalized odds, calibration, worst-group risk, and individual review may then be more important. The experiment intentionally shows that FairErase-FCL improves one predeclared disparity while reducing worst-group accuracy. A responsible paper should report that trade-off rather than describing the model as generally fair.

7.3 Privacy-fairness tension

Fairness estimation commonly needs sensitive attributes. Federated learning aims to keep such attributes local. Potential technical approaches include local group-statistic computation, secure aggregation, and privacy-preserving function tracking [3,4]. The present experiment directly uses synthetic group labels and implements none of those protections. A real deployment must not upload sensitive attributes to an ordinary coordination server merely to compute a fairness dashboard.

7.4 Educational governance implications

The European Union AI Act classifies several systems that influence access to education, placement, learning outcomes, or exam monitoring as high risk [12]. NIST AI RMF places fairness, privacy enhancement, transparency, reliability, and accountability within continuous lifecycle governance [13]. An unlearning record should therefore contain more than a binary deletion-success flag. It should identify the deletion unit, affected groups and tasks, metric changes, human approval, and available remedies.

8. Limitations

The benchmark uses binary synthetic groups and a binary classification task. It does not represent intersectional identities, continuous attributes, multilabel outcomes, or the institutional meaning of protected status. The group composition and label mechanism are selected by the researcher. The fairness penalty is a smooth proxy for equal opportunity, while the final decisions remain thresholded. The client-removal step is based on gradient influence and does not provide a universal formal unlearning guarantee. The study also does not evaluate an adversary who uses deletion requests to manipulate group outcomes, a concern raised by adaptive and attack-aware unlearning research [14].

9. Conclusion

FairErase-FCL extends fairness evaluation from initial training to the post-deletion release decision. In the seeded synthetic experiment, the method reduces the equal-opportunity gap from 8.05 to 6.65 percentage points at 24% of full-retraining computation, but loses approximately 0.42 points of balanced accuracy and does not improve worst-group accuracy. This is not evidence that the model is universally fair. It is evidence that deletion efficacy, average utility, and group distribution should be reported together and governed by an explicit choice of fairness objective.

Data and code availability

No real educational or demographic records are used. The synthetic generator, seeds, per-run results, and plotting code are included in the coursework directory.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *AISTATS*, PMLR 54, 1273-1282. [source]

[2] Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2020). Fair Resource Allocation in Federated Learning. *ICLR*. [source]

[3] Rychener, Y., Kuhn, D., and Hu, Y. (2025). Global Group Fairness in Federated Learning via Function Tracking. *AISTATS*, PMLR 258, 865-873. [source]

[4] Pentyala, S., et al. (2025). Privacy-Preserving Group Fairness in Cross-Device Federated Learning. *FAccT*. [source]

[5] Karimireddy, S. P., Kale, S., Mohri, M., et al. (2020). SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. *ICML*, PMLR 119, 5132-5143. [source]

[6] Hardt, M., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *NeurIPS 29*. [source]

[7] Liu, G., Ma, X., Yang, Y., Wang, C., and Liu, J. (2020). Federated Unlearning. arXiv:2012.13891. [source]

[8] Fraboni, Y., Van Waerebeke, M., Scaman, K., et al. (2024). SIFU: Sequential Informed Federated Unlearning for Efficient and Provable Client Unlearning in Federated Optimization. *AISTATS*, PMLR 238, 3457-3465. [source]

[9] Nguyen, T.-H., Vu, H.-P., Nguyen, D. T., et al. (2024). Empirical Study of Federated Unlearning: Efficiency and Effectiveness. *ACML*, PMLR 222, 959-974. [source]

[10] Özdenizci, O., Rueckert, E., and Legenstein, R. (2025). Privacy-Aware Lifelong Learning. arXiv:2505.10941. [source]

[11] Serra, J., Suris, D., Miron, M., and Karatzoglou, A. (2018). Overcoming Catastrophic Forgetting with Hard Attention to the Task. *ICML*, PMLR 80, 4548-4557. [source]

[12] European Parliament and Council. (2024). Regulation (EU) 2024/1689: Artificial Intelligence Act, Recital 56 and Annex III. [source]

[13] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. [source]

[14] Gupta, V., Jung, C., Neel, S., et al. (2021). Adaptive Machine Unlearning. *NeurIPS 34*. [source]

[S1] Tao, J., Liu, Z., Lyu, R., and Cao, X. (2026). A Scalable Data Governance Architecture for Privacy-Aware Intelligent Learning Systems in Lifelong. Course-supplied PDF manuscript.