Journal Pre-proof

Adaptive feature unlearning for trustworthy medical imaging privacy

Zhongyi Han, Bin Wang, Shenjing Wu, Juexiao Zhou, Gongning Luo, Benzheng Wei, Xin Gao

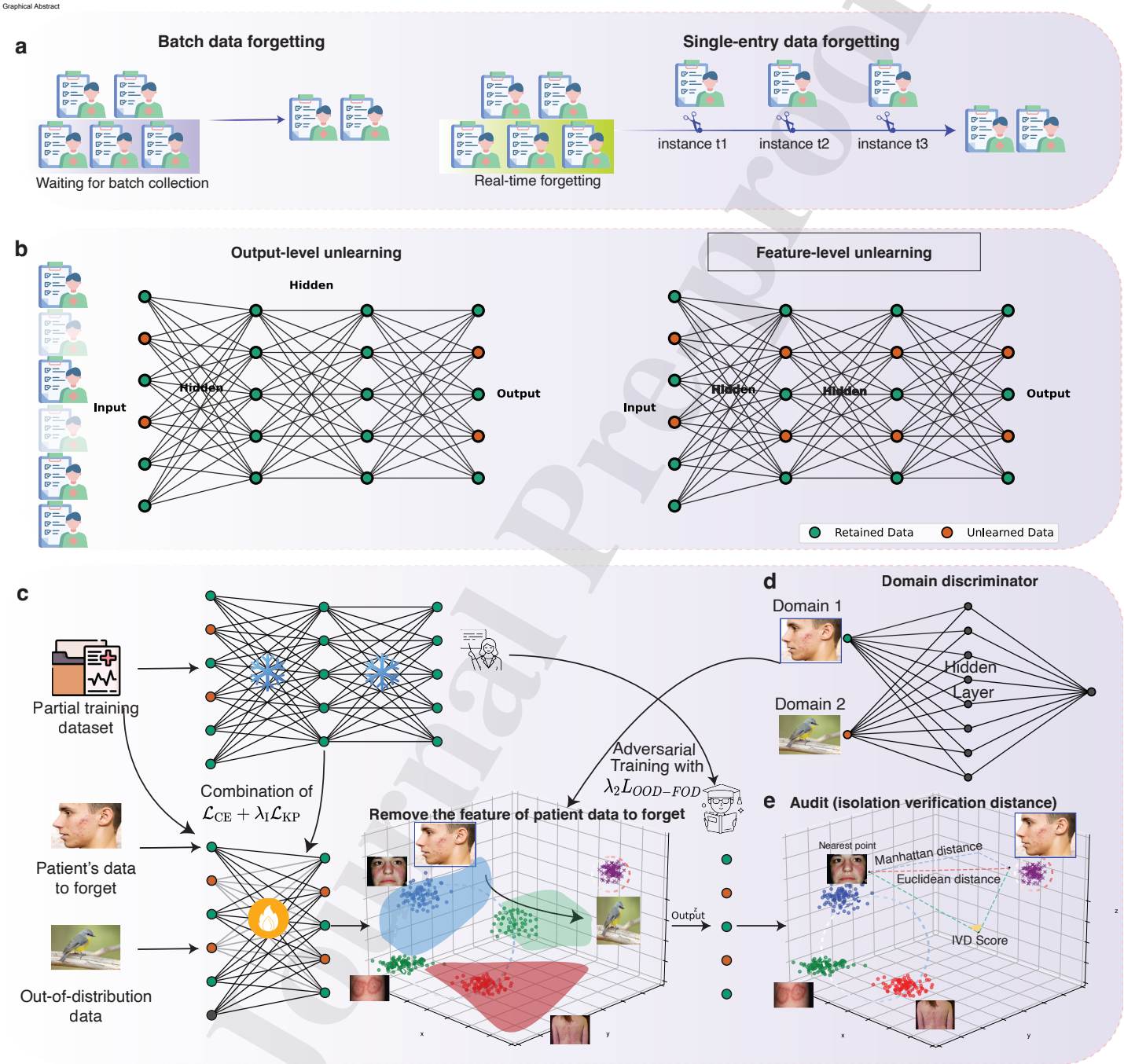
PII: S1361-8415(26)00220-3
DOI: https://doi.org/10.1016/j.media.2026.104151
Reference: MEDIMA 104151

To appear in: *Medical Image Analysis*

Received date : 2 October 2025
Revised date : 22 April 2026
Accepted date : 31 May 2026

Please cite this article as: Z. Han, B. Wang, S. Wu et al., Adaptive feature unlearning for trustworthy medical imaging privacy. *Medical Image Analysis* (2026), doi: https://doi.org/10.1016/j.media.2026.104151.

Graphical Abstract



Revised manuscript (marked up)

Adaptive Feature Unlearning for Trustworthy Medical Imaging Privacy

Zhongyi Han[a,b,*], Bin Wang[c,*], Shenjing Wu[c,*], Juexiao Zhou[b], Gongning Luo[b], Benzheng Wei[c,d,**], Xin Gao[b,**]

[a]School of Software, Shandong University, Jinan, 250100, China
[b]Computer Science Program, CEMSE Division; Center of Excellence on Smart Health; Center of Excellence for Generative AI, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Kingdom of Saudi Arabia
[c]Center for Medical Artificial Intelligence; Qingdao Academy of Chinese Medical Sciences; Institute of Marine Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Qingdao, 266112, China
[d]School of Medical Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan, 250355, China

Abstract

Deep learning has become integral to medical imaging, but its tendency to memorize training data poses serious risks for patient privacy. Machine unlearning offers a potential remedy by revoking sensitive information, yet existing approaches face three key limitations: (1) they often achieve only output-level changes while residual feature representations remain; (2) they rely on batch retraining, making real-time removal of individual patient images infeasible; and (3) they lack rigorous metrics to verify forgetting in feature space. We propose *AdaptForget*, a domain-adaptive feature-level unlearning framework for privacy-preserving medical image analysis. AdaptForget introduces out-of-distribution (OOD) guidance to disentangle forgotten data from retained data in the feature manifold, supported by a theoretical feature-level unlearning bound. To prevent feature collapse, we design an OOD-driven feature–output disentanglement loss that enforces structured removal of forgotten data. To enable timely revocation, we formalize the task of single-entry forgetting, allowing immediate erasure of individual patient records. For objective auditing, we propose the isolation verification distance, a novel metric that quantifies feature separation and provides interpretable evidence of forgetting. Extensive experiments on four medical imaging benchmarks (histopathology, retinal fundus, dermatology, and OCT) as well as complementary healthcare record datasets demonstrate that AdaptForget achieves state-of-the-art privacy protection while preserving model utility. Code is publicly available at `https://github.com/wangbrav/AdaptForget`.

Keywords: Healthcare, Single-entry Data Forgetting, Feature-level Unlearning, Machine Unlearning.

1. Introduction

Revoking personal private data is a fundamental human right, especially in healthcare where patient privacy is paramount. The increasing digitalization of the healthcare sector has made it more vulnerable to cybersecurity incidents, exposing sensitive patient data to heightened risks of cyberattacks (ENISA, 2023; Lampropoulos, 2023). Data breaches are escalating both in frequency and severity; from 2015 to 2022, 32% of all recorded breaches occurred in healthcare, leading to the most costly incidents over the past 13 years (Security Intelligence, 2023). This alarming trend has been intensified by the COVID-19 pandemic, with breached records surging to 133 million in 2023 (Muthuppalaniappan, 2021; HIPAA Journal, 2023). In 2024, the largest breach affected over 11 million individuals, marking it as the second-largest healthcare data breach on record (HIPAA Journal, 2023). Many EU countries are investing substantially in advanced e-health tools and applications to enhance healthcare services (Greer, 2022). Significant challenges remain despite mitigation efforts and legislation like the European Union's GDPR and the United States' HIPAA advocating for "the right to be forgotten" (Voigt, 2017; U.S. Congress, 1996). The core challenge is that these measures alone are insufficient to fully safeguard patient privacy against modern technologies like deep learning due to the memorization ability.

Deep learning is transforming the healthcare sector (Alowais, 2023), with Google's AI systems for detecting diabetic retinopathy (Gulshan, 2016) and predicting acute kidney injury (Tomašev, 2019). However, deep neural networks' ability to memorize training data makes them susceptible to leaking patient sensitive information. Various attacks (Shokri, 2017; Fredrikson, 2015) can recover sensitive patient information from deep neural networks, including but not limited to mental health conditions, substance abuse histories, rare diseases, identifiable facial features, unique anatomical features, genetic predispositions, and other confidential medical details. Machine unlearning, which aims to delete specific data records from the memory of trained deep learning models without the need for full retraining (Zhou, 2023; Bourtoule, 2021; Nguyen, 2022), has emerged as a potential solution to these privacy concerns. Although a variety of model-agnostic (Bourtoule, 2021; Cha, 2024), model-specific (Baumhauer, 2022; Schelter, 2021), and data-centric machine unlearning techniques (Bourtoule, 2021; Tarun, 2024; Graves, 2021) have been proposed, the lack of completeness, timeliness, and objectivity, has prevented ma-

*These authors contributed equally to this work.
**Corresponding author.
Email addresses: `wbz99@sina.cn` (Benzheng Wei), `xin.gao@kaust.edu.sa` (Xin Gao)

chine unlearning algorithms from robust protection of healthcare data.

Systematic analyses (Zhou, 2023; Nguyen, 2022; Qu, 2023) of state-of-the-art (SOTA) machine unlearning in healthcare over the past 10 years highlight three bottleneck issues hindering meaningful progress: incomplete forgetting, lack of real-time forgetting capabilities, and poor metrics for evaluating forgetting completeness. As a result, the healthcare industry often resorts to retraining models from scratch after deleting the records that must be forgotten to ensure complete data removal (Nguyen, 2022; Bourtoule, 2021; Graves, 2021). While current SOTA machine unlearning algorithms in healthcare often report high revocation success rates (e.g., p-value < 0.01) against ensembled membership attacks (EMA) (Huang, 2021), these success rates primarily reflect performance on output-level forgetting, where the output distributions of the forgotten data differ from those of the retained data (Tarun, 2023; Cha, 2024; Foster, 2024; Zhou, 2023). However, when evaluated for the degree of disentanglement in latent feature space, about 99% of the forgotten points remain entangled with the retained points (see Appendix A for a proof of feature entanglement). This leaves room for various feature-level attack techniques to recover patient privacy information, such as feature inversion attacks (Dosovitskiy, 2016; Mahendran, 2015), gradient leakage attacks (Zhu, 2019), attribute inference attacks (Ganju, 2018), and knowledge distillation attacks (He, 2019).

Although batch data forgetting is commonly studied, it is often necessary to forget single-entry data individually to enable real-time revocation of patient data without waiting for a batch to accumulate. Accordingly, we propose the new concept of single-entry data forgetting that allows immediate compliance with data removal requests. It is crucial in dynamic healthcare environments where individual patient data may need to be withdrawn at any time, particularly those related to rare diseases, sensitive conditions, or histories of substance abuse. The illustration comparison between batch data forgetting and single-entry data forgetting can be seen in Figure 1a (see Appendix B for a full discussion). However, revocation of single entry data is more difficult (as current SOTA algorithms report p-value $= 1$ (Zhou, 2023)) due to the complex interdependencies between data points in the manifold space (Serra, 2018) and the lack of precise audit metrics (Shokri, 2017; Zhou, 2023). Common audit metrics are accuracy, F1 score, and EMA (calculating the p-value of correctness, confidence, and entropy) as primary metrics (Zhou, 2023). Since these metrics evaluate the forgetting degree at the output space, improvements in these metrics do not necessarily translate into improvements in the feature space (see Appendix C for a summary of metric issues).

To address the above-mentioned three key issues of incomplete forgetting, limited timeliness, and lack of auditability simultaneously, we present AdaptForget: a medically motivated framework with a modality-independent, transferable mechanism for domain-adaptive feature-level machine unlearning in vulnerable healthcare applications. AdaptForget uses a teacher-student framework where a teacher model trained on the full dataset distills knowledge to a newly initialized student model that is trained on a subset of the training with limited iterations. To enable AdaptForget to achieve complete forgetting, we employ an out-of-distribution (OOD) data-driven feature-output disentanglement mechanism to revoke the forgotten data at the feature level and the output level. The key innovation is that we introduce OOD data to learning the domain-invariant representations, allowing the model to disentangle the forgotten data from the retained data at the feature level (Figure 1b). The input to AdaptForget consists of the retaining data, the forgotten data, and the OOD data that can represent and control the unlearning direction in high-dimensional feature space, as shown in Figure 1c. This design not only enables complete forgetting of groups of data but, more importantly, allows for single-entry data forgetting in real time because the complex interdependencies between a single point and others are broken such that each patient record is fully removed from the model's learned representations. Finally, we propose a new audit metric to address poor metrics issues. This metric quantitatively assesses the completeness of forgetting at the feature level by measuring the feature distance between the forgotten data and its nearest neighbors. AdaptForget is fundamentally shaped by three clinical deployment constraints (feature-level erasure obligations, multi-center OOD distribution shift, and prohibitive retraining costs) that co-occur most severely in medical AI.

We summarize our principal contributions as follows.

- To the best of our knowledge, we are the first to rigorously define and analyze the challenges associated with single-entry data forgetting in healthcare.
- We propose the domain-adaptive feature-level unlearning framework and OOD-diven feature-output disentanglement loss to enforce structured unlearning, ensuring that forgotten data is irreversibly erased from both feature representations and model predictions.
- We propose the first feature-level audit metric that quantitatively evaluates the extent of feature disentanglement by measuring manifold distance between forgotten data and nearest retained counterparts.
- We validate AdaptForget on seven diverse healthcare benchmarks, demonstrating state-of-the-art performance in both batch and single-entry forgetting.

2. Related Work

2.1. Machine Unlearning

Machine unlearning has emerged as a critical research area that aims to enable deep learning models to erase specific data records while preserving overall performance. Existing approaches can be broadly classified into three categories: model-agnostic, model-specific, and data-centric methods (Nguyen, 2023). Model-agnostic techniques, such as Instance-wise Unlearning (Cha, 2024), FisherForgetting (Golatkar, 2020), AFS (Zhou, 2023) and SSD-Tuning (Foster, 2024), apply general unlearning mechanisms that are independent of model architecture, typically using weight perturbation, adversarial instance training, or selective parameter dampening. Model-specific methods leverage the unique properties of particular

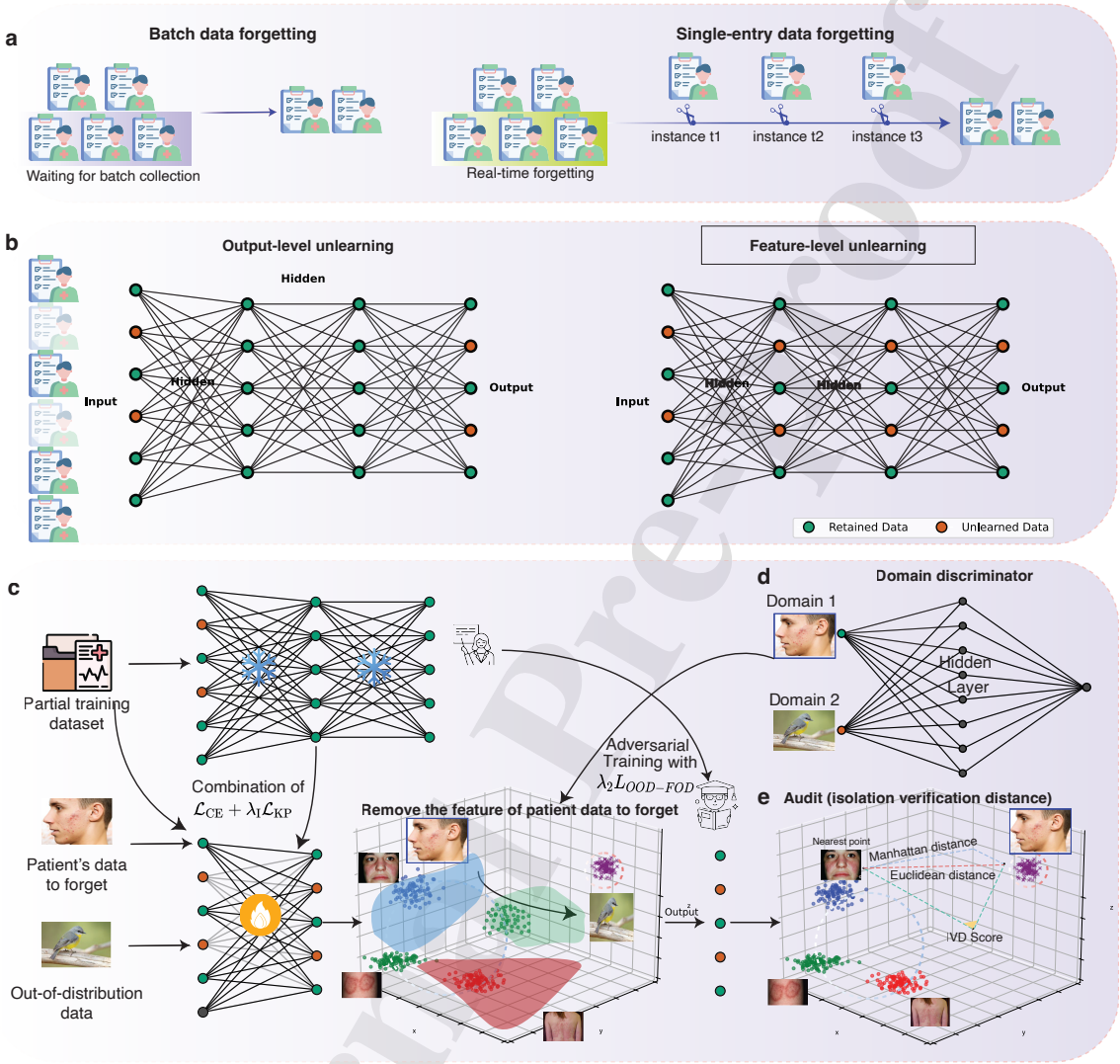


Figure 1: An overview of the AdaptForget framework that improves completeness, timeliness, and objectivity through real-time forgetting and feature-level unlearning. **a** Comparison between batch data forgetting and single-entry data forgetting. Batch data forgetting needs waiting for batch collection, in contrast, single-entry data forgetting enables real-time forgetting of forgotten data one-by-one. **b** Comparison between output-level unlearning and feature level unlearning. Output-level unlearning aims to make the output distributions of forgotten data differ from the retained data, while feature-level unlearning can remove the complete information in the feature space. **c** The overall architecture of the AdaptForget framework. The system leverages a teacher-student model paradigm. The teacher model, trained on the full dataset, transfers knowledge to the student model. **d** The overall representation of domain discriminator enabling that forgotten data points are effectively shifted towards the out-of-distribution (OOD) data. **e** Audit process using isolation verification distance, measuring manifold distance between forgotten data and nearest retained counterparts.

architectures, such as pruning decision trees (Schelter, 2021) or unlearning via final-layer reconfiguration (Goel, 2022). Data-centric approaches, including SISA (Bourtoule, 2021) and Amnesiac Unlearning (Graves, 2021), manipulate training data before or during the learning process to facilitate more efficient forgetting. Although these methods have shown effectiveness in controlled environments, they focus mainly on output level unlearning by modifying prediction distributions rather than achieving deeper structural modifications in the model feature space.

Our work addresses key limitations of existing unlearning methods in healthcare by shifting from output-level modifica-

tions to feature-level disentanglement. Traditional approaches leave sensitive patient data embedded in feature representations, making them vulnerable to inversion and leakage attacks (Dosovitskiy, 2016; Zhu, 2019; Ganju, 2018). To resolve this, we introduce OOD-driven domain adversarial training to enforce domain-invariant feature unlearning, ensuring complete disentanglement of forgotten data. Unlike batch-based methods that delay compliance with privacy regulations, AdaptForget enables real-time single-entry forgetting, allowing immediate removal of individual records without compromising model integrity. Furthermore, existing audit mechanisms lack feature-space verification. To address this, we propose the isolation verification distance, which quantifies the separation of forgotten and retained data in the feature space, offering an objective assessment of the completeness of unlearning. These innovations establish AdaptForget as a principled and practical solution for privacy-preserving machine unlearning in healthcare.

2.2. OOD-based Learning

Out-of-Distribution (OOD) learning refers to a class of machine learning techniques that aim to generalize beyond the training distribution, enabling models to detect, adapt to, or generalize across data that significantly deviates from the original training set. OOD learning has been extensively studied in two major directions: OOD detection and OOD generalization (Lu, 2024; Zhao, 2023; Han, 2022a). OOD detection focuses on identifying whether a given input falls outside the training distribution, which is crucial in applications such as medical diagnosis, autonomous driving, and security-sensitive systems (Yang, 2024). Traditional approaches include classification methods that design a scoring function to measure the uncertainty of test data (He, 2024; Gomes, 2022), and density-based methods that model the ID distribution using probabilistic models and consider test data in low-density regions as OOD data (Jiang, 2022; Zhang, 2021; He, 2022; Gomes, 2022). OOD generalization, on the other hand, aims to enhance model robustness when encountering data distributions that differ from the training environment (Han, 2025). Domain generalization (DG) (Zhou, 2022; Robey, 2021) and domain adaptation (DA) (Han, 2022b; Ganin, 2016; Han, 2023) are two key strategies in this area. DG methods attempt to learn invariant representations that can generalize to unseen domains without requiring labeled target data, often using adversarial training (Li, 2018b) or meta-learning techniques (Li, 2018a). DA methods leverage labeled source domain data to adapt to an unlabeled target domain, commonly through domain adversarial training (Ganin, 2015), self-training (Liu, 2021), or contrastive learning (Thota, 2021). In healthcare, OOD generalization techniques are widely adopted to improve model robustness across different patient demographics, imaging devices, and clinical settings (Guan, 2021; Li, 2020).

Despite progress in OOD-based learning, our work is the first to introduce OOD-based learning into machine unlearning, addressing a crucial gap in privacy-preserving machine learning. Unlike prior OOD detection methods that identify anomalous inputs or OOD generalization techniques that aim to improve model robustness, we leverage OOD learning as a mechanism for feature-level unlearning, using DA strategies to disentangle forgotten data from retained data explicitly.

3. Methodology

Design rationale under medical deployment constraints.. AdaptForget is designed around three clinical deployment constraints that are especially acute in medical AI. *First*, patient-level removal under HIPAA/GDPR requires erasure beyond output behavior, since residual feature representations can still be exploited via feature inversion or attribute inference attacks, directly motivating the OOD-FOD loss for feature-level forgetting. *Second*, clinical data are routinely collected across institutions with heterogeneous scanners, acquisition protocols, and patient populations, so unlearning must preserve diagnostic generalization under multi-center OOD shift, motivating structured OOD selection rather than arbitrary noise. *Third*, full model retraining in regulated clinical environments incurs repeated ethical approval and data governance overhead, making post-hoc, single-forward-pass deletion efficiency essential, which motivates the teacher-student update mechanism. Together, these three constraints shaped the key design choices and methodology in AdaptForget.

We begin by introducing the learning set-up, defining batch and single-entry data forgetting, and comparing output-level and feature-level unlearning (Section 3.1). Next, we present the overall framework of AdaptForget (Section 3.2). We then formalize OOD-guided feature-level unlearning, establishing a unlearning generalization bound and proposing the OOD-driven feature-output disentanglement loss to enforce structured unlearning (Section 3.3). Finally, we introduce the new feature-level audit metric that quantitatively assesses the extent of forgetting in latent space (Section 3.5).

3.1. Learning Set-Up

In this subsection, we formalize the learning set-up for machine unlearning. We first define batch data forgetting and its single-entry special case, then introduce OOD data, and finally distinguish output-level unlearning from feature-level unlearning. These definitions establish the problem formulation and terminology used throughout the methodology.

Definition 1 (Batch Data Forgetting)**.** *Given a trained deep learning model* $h_\Theta : \mathcal{X} \to \mathcal{Y}$ *with parameters* Θ*, let* $\mathrm{QO} = \{(x^i, y^i)\}_{i=1}^{n_t}$ *denote the training dataset and let* $\mathrm{QF} = \{(x^i, y^i)\}_{i=1}^{n_f} \subset \mathrm{QO}$ *denote the batch to be removed. Let* $h_{\Theta\setminus\mathrm{QF}}$ *be the model retrained from scratch on* $\mathrm{QO} \setminus \mathrm{QF}$*. To disambiguate notation, let* $r_\Theta(x)$ *denote an internal representation, and define the joint feature–output variable* $Z_\Theta(x) := \big(r_\Theta(x),\, h_\Theta(x)\big)$*. A machine unlearning algorithm achieves* ***batch data forgetting*** *if it produces* $h_{\Theta'}$ *such that the joint distribution of representations and predictions on* QF *under* $h_{\Theta'}$ *is statistically indistinguishable from that of the retrained reference, while retaining predictive performance on* $\mathrm{QO} \setminus \mathrm{QF}$*. Formally,*

$$d\Big(\mathsf{Law}\big(Z_{\Theta'}(X)\big)\big|_{X\sim\hat{P}_{\mathrm{QF}}},\ \mathsf{Law}\big(Z_{\Theta\setminus\mathrm{QF}}(X)\big)\big|_{X\sim\hat{P}_{\mathrm{QF}}}\Big) \ \le\ \varepsilon. \tag{1}$$

Table 1: Summary of Key Notations

Symbol	Description
Spaces & Data	
$\mathcal{X},\ \mathcal{Y}$	Input space and label space
(x^i, y^i)	The i-th training sample (input, label)
QO	Training dataset / Query dataset overlapped with training set $\{(x^i, y^i)\}_{i=1}^{n_t}$
QF	Forget set $\{(x^i, y^i)\}_{i=1}^{n_f} \subset$ QO
QNO	Never-seen holdout set
OOD	Out-of-distribution dataset drawn from Q
Distributions	
$P_r,\ P_f,\ Q$	Distributions of retained, forgotten, and OOD data
$\hat{P}$	Empirical distribution over a finite sample
$\mathsf{Law}(\cdot)$	Induced distribution (law) of a random variable
$d(\cdot,\cdot)$	Statistical distance between two distributions
Models & Parameters	
h_Θ	Trained model with parameters Θ
$h_{\Theta'}$	Updated model after unlearning
$h_{\Theta\setminus\mathrm{QF}}$	Model retrained from scratch without QF
F_θ	Feature extractor parameterized by θ
C_ϕ	Classifier parameterized by ϕ
D_ψ	Domain discriminator parameterized by ψ
$r_\Theta(x)$	Internal feature representation of input x
$Z_\Theta(x)$	Joint feature–output variable $(r_\Theta(x),\ h_\Theta(x))$
Loss Functions & Hyperparameters	
$\mathcal{L}$	Total training loss
$\mathcal{L}_{\mathrm{CE}}$	Cross-entropy loss for task classification
$\mathcal{L}_{\mathrm{KP}}$	Knowledge purification loss
$\mathcal{L}_{\text{OOD-FOD}}$	OOD-driven feature-output disentanglement loss
$\lambda_1,\ \lambda_2$	Hyperparameters balancing loss terms
Theoretical Quantities	
$\mathcal{H}$	Hypothesis class of real-valued functions
$\mathfrak{R}_n(\mathcal{H})$	Rademacher complexity for n samples
$d_{\mathcal{H}\Delta\mathcal{H}}$	$\mathcal{H}\Delta\mathcal{H}$-discrepancy between distributions
$\hat{\epsilon}_S(h)$	Empirical 0–1 risk on sample S
λ	Alignment error (ideal joint error of P_r and P_f)
$\varepsilon,\ \delta$	Tolerances for statistical-distance and utility constraints
$\varepsilon_o,\ \varepsilon_f$	Tolerances for output-level and feature-level unlearning

In addition, we require utility preservation on the retained set: $R_{\mathrm{QO}\setminus\mathrm{QF}}(h_{\Theta'}) - R_{\mathrm{QO}\setminus\mathrm{QF}}(h_\Theta) \le \delta$. Here $\mathsf{Law}(\cdot)$ denotes the induced distribution (law) of a random variable, and $d(\cdot,\cdot)$ is a statistical distance, and $\varepsilon > 0$, $\delta \ge 0$ are given tolerances for the statistical-distance constraint and the utility-preservation constraint, respectively.

Definition 2 (Single-Entry Data Forgetting)**.** *Single-entry data forgetting is a special case of batch data forgetting where $n_f = 1$. Given a trained deep learning model $h_\Theta : \mathcal{X} \to \mathcal{Y}$, a training dataset QO $= \{(x^i, y^i)\}_{i=1}^{n_t}$ sampled from $P(x)$, and a single sample (x_f, y_f) to be forgotten, a machine unlearning algorithm achieves **single-entry data forgetting** if it produces an updated model $h_{\Theta'}$ satisfying:*

$$h_{\Theta'}(x_f) \approx h_{\Theta_{\setminus\{x_f, y_f\}}}(x_f)\,, \tag{2}$$

where $h_{\Theta_{\setminus\{x_f, y_f\}}}$ denotes a model retrained without (x_f, y_f). Additionally, for effective real-time forgetting, the unlearning process should minimize computational overhead, ensuring the model remains usable in dynamic environments.

Definition 3 (Out-of-Distribution (OOD) Data)**.** *Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space. Given a training dataset QO $= \{(x^i, y^i)\}_{i=1}^{n_t}$ sampled from an in-distribution $P(x, y)$, we define an OOD dataset as a set of instances OOD $= \{(x^i, y^i)\}_{i=1}^{n_d}$ drawn from a different distribution $Q(x, y)$, where $P(x) \neq Q(x)$. The OOD dataset is used to aid domain adaptation in machine unlearning by aligning its representations with the forgotten data distribution.*

Definition 4 (Output-Level Unlearning)**.** *Let $h_\Theta : \mathcal{X} \to \mathcal{Y}$ be a trained model. Given a training set* QO *and a forgotten subset* QF $\subset$ QO*, we say $h_{\Theta'}$ achieves output-level unlearning if the* output distribution *on* QF *matches that on a never-seen holdout* QNO*:*

$$d\Big(\hat{P}_{h_{\Theta'}(X)\,|\,X\in\mathrm{QF}},\ \hat{P}_{h_{\Theta'}(X)\,|\,X\in\mathrm{QNO}}\Big) \le \varepsilon_o, \tag{3}$$

here $d(\cdot,\cdot)$ denotes a statistical distance, and $\hat{P}_{f(X)\,|\,X\in\mathcal{S}}$ denotes the empirical distribution of $\{f(x^i)\}$ induced by uniformly sampling x from a finite set $\mathcal{S}$.

Definition 5 (Feature-Level Unlearning)**.** *Let $r_\Theta(x)$ denote an internal feature representation (e.g., the penultimate-layer embedding). A model $h_{\Theta'}$ achieves feature-level unlearning if the* feature distribution *of forgotten samples becomes indistinguishable from that of unseen data:*

$$d\Big(\hat{P}_{r_{\Theta'}(X)\,|\,X\in\mathrm{QF}},\ \hat{P}_{r_{\Theta'}(X)\,|\,X\in\mathrm{QNO}}\Big) \le \varepsilon_f. \tag{4}$$

3.2. The Framework of AdaptForget

Machine unlearning becomes fundamentally more challenging when the forget set collapses to a single entry ($|D_f| = 1$), as the nature of the optimization problem changes qualitatively rather than merely scaling down. Three core challenges emerge in this regime: (1) Signal Dilutio: the forgetting gradient contributes only $O(1/|D_r|)$ to the total gradient, causing it to be overwhelmed by the retain set; (2) Ill-Posed Forgetting Target: a single data point cannot define a meaningful distribution to align with, making standard divergence-based objectives undefined; (3) Optimization Instability: single-sample gradients exhibit high variance and are susceptible to target drift during iterative updates.

To achieve complete and verifiable unlearning, we introduce AdaptForget, a domain-adaptive machine unlearning framework designed to ensure that forgotten data is thoroughly removed from both the model's output predictions and its internal feature representations. As shown in Figure 1c, AdaptForget adopts a teacher-student learning paradigm, where a student model is trained under the supervision of a pretrained teacher model. However, instead of directly inheriting knowledge from the teacher model, the student model undergoes a controlled adaptation process that explicitly removes any influence of the forgotten data.

Given an initial pre-trained teacher model h_{teacher}, the student model h_{student} is iteratively trained using a subset of the training dataset while enforcing the unlearning of specific query data points. The student model consists of the following components: 1) Feature Extractor F_θ: A deep network parameterized by θ that learns feature representations of input data. 2) Classifier C_ϕ: A model head parameterized by ϕ that predicts output labels based on extracted features. 3) Domain Discriminator

D_ψ: A neural network parameterized by ψ trained to distinguish between retained data and forgotten data, playing a key role in ensuring feature-level unlearning.

The student model undergoes adversarial training, where the feature extractor learns to remove forgotten data dependencies while maintaining task performance on the retained data. The AdaptForget framework is trained by minimizing a multi-objective loss function, which balances classification performance, knowledge transfer, and structured feature-output unlearning:

$$\min_{\theta,\phi} \max_{\psi} \mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{KP}} + \lambda_2 \mathcal{L}_{\text{OOD-FOD}} \,. \tag{5}$$

where $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss for task classification, ensuring that the model maintains predictive performance on retained data. The knowledge purification loss $\mathcal{L}_{\text{KP}}$ refines the teacher-student knowledge transfer by eliminating residual dependencies on forgotten data (Zhou, 2023). The newly introduced OOD-Driven Feature-Output Disentanglement Loss $\mathcal{L}_{\text{OOD-FOD}}$ enforces feature-space alignment between forgotten data and OOD data while simultaneously ensuring output-level prediction consistency. The hyperparameters λ_1 and λ_2 regulate the relative contributions of knowledge purification and structured unlearning. Minimization for the feature extractor parameters θ and classifier parameters ϕ reduces the overall loss, while the maximization for the domain discriminator parameters ψ ensures that the discriminator can effectively distinguish between forgotten data and OOD data. The formulation of $\mathcal{L}_{\text{OOD-FOD}}$ is presented below (Section 3.3), while the complete workflow of AdaptForget is detailed in Algorithm 1.

Algorithm 1 AdaptForget

Require: Pre-trained teacher model h_{teacher}
Require: Student model h_{student} with parameters θ, ϕ, ψ
Require: Query dataset QF $= \{(x^i, y^i)\}_{i=1}^{n_f}$ for forgetting
Require: OOD dataset OOD $= \{(x^i, y^i)\}_{i=1}^{n_d}$
Require: Training dataset QO $= \{(x^i, y^i)\}_{i=1}^{n_t}$
Require: Learning rate η, weight decay coefficient γ, balancing parameters λ_1, λ_2, and λ_3
Ensure: Updated student model parameters θ, ϕ, ψ

procedure ADAPTFORGET
  while not converged **do**
    Compute the total loss: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{KP}} + \lambda_2 \mathcal{L}_{\text{OOD-FOD}}$.
    Update Student Model Parameters:
    $\theta(t+1) = \theta(t) - \eta \left(\frac{\partial \mathcal{L}}{\partial \theta(t)} + \gamma \theta(t) \right)$
    $\psi(t+1) = \psi(t) - \eta \left(\frac{\partial \mathcal{L}_{\text{OOD-FOD}}}{\partial \psi(t)} + \gamma \psi(t) \right)$
    $\phi(t+1) = \phi(t) - \eta \left(\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \phi(t)} + \gamma \phi(t) \right)$
  return Updated student model parameters

3.3. OOD-Guided Feature-Level Unlearning

Adaptforget directly addresses the ill-posed forgetting target challenge by providing a well-defined proxy distribution: instead of attempting to estimate a distribution from a single forgotten sample, we align it with a carefully selected OOD dataset that serves as a stable reference. Specifically, We introduce OOD-guided feature-level unlearning, which explicitly forces the feature representations of forgotten data to be aligned with out-of-distribution (OOD) data. Feature-space forgetting must be carefully controlled to avoid performance degradation. A naive approach such as directly disturbing the feature representations may lead to an uncontrolled shift in the model's learned feature space, resulting in significant feature space collapse. OOD-guided unlearning provides a structured forgetting direction by aligning the forgotten data's features with those of OOD data, thereby preserving model stability while ensuring that forgotten data is effectively erased from the learned representations.

The selection of OOD data plays a critical role in the effectiveness of this approach. Ideally, the OOD data should be sufficiently distinct from the training data to facilitate feature separation, while remaining within a controlled distribution shift to prevent excessive deviation from the original feature space. We provide a rigorous empirical generalization bound for OOD-guided feature-level unlearning, where one aims to remove the influence of a forgotten dataset by aligning it with an OOD dataset, all while preserving performance on a retained dataset. Formal definitions of the empirical $\mathcal{H}\Delta\mathcal{H}$-discrepancy, empirical Rademacher complexity, and alignment error λ are provided in Appendix D; we present the main proposition directly below.

Proposition 1 (Empirical Feature-Level Unlearning Bound)**.** *Consider the batch data forgetting setting (Definition 1) with a hypothesis class $\mathcal{H}$ of real-valued functions, and let $\mathfrak{R}_n(\mathcal{H})$ denote its (expected) Rademacher complexity for n labeled samples. Suppose we have: QO $\sim P_r$, QF $\sim P_f$, OOD $\sim Q$, and a final hypothesis $h \in \mathcal{H}$ after OOD-guided unlearning. Let $\hat{\epsilon}_S(h)$ be the* empirical 0–1 risk *of h on the labeled sample QO of size n. Let $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_r, \hat{Q})$ and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_f, \hat{Q})$ be the* empirical *$\mathcal{H}\Delta\mathcal{H}$-discrepancies estimated from unlabeled samples for QO, OOD and QF, OOD. Then, with probability at least $1-\delta$ over the random draw of QO, the following inequality holds for every $h \in \mathcal{H}$:*

$$\begin{aligned} \epsilon_{P_r}(h) \;\le\; & \hat{\epsilon}_S(h) \;+\; \left| \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_r, \hat{Q}) \;-\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_f, \hat{Q}) \right| \\ & + \lambda \;+\; 2\,\mathfrak{R}_n(\mathcal{H}) \;+\; 2\,\mathfrak{R}_m(\mathcal{H}) \\ & + \sqrt{\frac{\log(2/\delta)}{2n}} \;+\; \sqrt{\frac{\log(2/\delta)}{2m}}. \end{aligned} \tag{6}$$

See Appendix D for the proof.

Theoretical Contributions. While Proposition 1 employs standard theoretical tools (Rademacher complexity and $\mathcal{H}\Delta\mathcal{H}$-discrepancy), its key contribution lies in the *problem formulation* rather than the proof techniques. Unlike standard domain adaptation theory, which addresses two-domain scenarios (source and target), Proposition 1 introduces a three-distribution framework (retained P_r, forgotten P_f, and OOD Q), formalizing the mechanism of forgetting through OOD alignment at the feature level. This formulation is unique to

feature-level unlearning and provides an actionable criterion for OOD selection: the bound reveals that Q should be sufficiently separated from P_f to enable effective forgetting, while maintaining moderate distance from P_r to preserve model utility. This theoretical insight directly motivates our Structured OOD strategy (detailed below) and establishes a principled connection between the generalization bound and our algorithmic design (the OOD-FOD Loss). Importantly, the proposition addresses feature-level unlearning through the joint variable $Z_\Theta(x) = (r_\Theta(x), h_\Theta(x))$, filling a critical gap in existing unlearning theory, which predominantly provides output-level guarantees.

Remark 1 (Tradeoff in OOD Selection for Feature-Level Unlearning)**.** *Proposition 1 quantifies the tradeoff in selecting OOD data for effective unlearning. A more distinct OOD distribution Q enhances feature separation by increasing $d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q)$, facilitating the removal of forgotten data. However, if P_r and Q are too divergent, it not only enlarges $d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q)$, degrading model generalization, but also indirectly increases the alignment error λ (see Appendix D), since a poor alignment can inflate $\epsilon_{P_f}(h^*)$. This results in residual dependencies in forgotten data. Conversely, choosing an OOD distribution Q that is too close to P_f may fail to achieve sufficient feature separation, limiting the forgetting effectiveness. Therefore, optimal OOD selection must balance the tradeoff between maximizing $d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q)$ for feature separation and minimizing λ to ensure efficient and trustworthy unlearning.*

Inspired by Proposition 1, we propose a principled OOD selection strategy for effective feature-level unlearning while maintaining model stability. Instead of using arbitrarily unrelated OOD data, such as natural images (e.g., cats, dogs) for forgetting CT scans, we generate structured OOD data tailored to the specific modality. For medical images, we introduce controlled corruptions such as Gaussian noise, salt-and-pepper noise, or adversarial perturbations, ensuring that the OOD data retains low-level image characteristics while eliminating domain-specific semantic information. For tabular clinical data, we construct OOD samples by shifting feature distributions through controlled perturbations (e.g., resampling from altered marginal distributions, injecting synthetic anomalies), preserving overall statistical consistency while disrupting learned feature associations.

To achieve structured forgetting, we introduce the OOD-Driven Feature-Output Disentanglement (OOD-FOD) Loss, which enforces both feature-level and output-level alignment between forgotten data and OOD data. We denote by $z = F_\theta(x)$ the extracted feature representation, $D_\psi(z)$ the domain discriminator's probability of the sample being from the forgotten data, and $C_\phi(z)$ is the classifier's softmax output. The OOD-FOD loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{OOD-FOD}} = &- \mathbb{E}_{z\sim P_f}\big[\log D_\psi(z)\big] - \mathbb{E}_{z\sim Q}\big[\log(1 - D_\psi(z))\big] \\ &+ \lambda_3 \mathbb{E}_{z_f\sim P_f, z_q\sim Q}\big[\mathrm{KL}\big(C_\phi(z_f) \parallel C_\phi(z_q)\big)\big]. \quad (7)\end{aligned}$$

This loss consists of two parts. The first two terms represent the domain adversarial loss that enforces feature alignment between forgotten data and OOD data by training a domain discriminator D_ψ. The second part is the feature-output alignment loss that ensures the classifier outputs for forgotten data match those of OOD data, removing decision boundary dependencies. Minimizing $\mathcal{L}_{\text{OOD-FOD}}$ ensures that forgotten data is indistinguishable from OOD data in both feature space and prediction space. Furthermore, we establish that minimizing $\mathcal{L}_{\text{OOD-FOD}}$ guarantees both feature-level and output-level disentanglement.

This guarantees that forgotten data is irreversibly transformed into representations and predictions that align with OOD data. Unlike traditional adversarial training (Ganin, 2016), which primarily focuses on feature alignment, our loss enforces joint alignment of both feature and output spaces, ensuring that forgotten data is independent of decision boundaries. Moreover, while traditional adversarial training is designed for domain adaptation, our method specifically targets structured forgetting, making it well-suited for privacy-preserving machine unlearning. Furthermore, the output alignment guarantee directly satisfies the requirement of Definition 4, as it ensures that the predictive distribution of forgotten data aligns with that of OOD data, making it indistinguishable from an unseen test set (QNO). Under standard regularity conditions, the formal convergence analysis is given in Appendix G.

3.4. OOD Selection and σ Calibration

While the theoretical framework above establishes the necessity of OOD guidance for feature-level unlearning, selecting the appropriate OOD distribution in practice remains non-trivial. Proposition 1 indicates that effective OOD guidance requires a candidate distribution that is sufficiently separated from the forgotten set while remaining moderately aligned with the retained domain. The following remark provides a directional account of how the OOD intensity parameter σ governs this tradeoff; the formal statement and proof are given in Appendix D.

Remark 2 (Directional role of OOD intensity σ)**.** *Under a noise-scale parameterization Q_σ (e.g., Gaussian with standard deviation σ) satisfying the monotonicity condition $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_{\sigma_2}) \geq \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_{\sigma_1})$ for $\sigma_2 > \sigma_1$, the triangle inequality gives a lower bound on the displacement of forgotten representations:*

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^\sigma, P_f^0) \;\geq\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_\sigma) - \eta_\sigma,$$

where $\eta_\sigma := \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^\sigma, Q_\sigma)$ is the residual matching error. This shows that increasing σ tends to strengthen forgetting, but a larger discrepancy involving Q_σ can loosen the retained-data risk bound, so stronger forgetting may reduce utility. This is a conditional, directional account, not a certified privacy guarantee. Since the optimal σ cannot be determined analytically, we adopt the following three-step empirical selection strategy:

- Step 1 (Selection): Select a candidate OOD type from the standardized pool (e.g., Gaussian noise).
- Step 2 (Calibration): Perform a sensitivity analysis across a range of OOD intensities (e.g., varying noise scale σ or shift magnitude).

- Step 3 (Optimization): Identify the optimal configuration by evaluating the privacy-utility trade-off using Forget Accuracy, OOD AUROC, Test Accuracy, and F1-score.

This strategy provides a practical starting point for narrowing the search space; dataset-specific tuning remains necessary. We validate this OOD selection strategy through multi-configuration analysis on the Camelyon17-WILDS dataset (see the validation results in the experiments section).

3.5. Isolation Verification Distance

According to Definition 5, the fundamental criterion for feature-level unlearning is that the feature representation of a forgotten sample should be statistically independent from its original representation before unlearning. However, directly verifying this statistical independence is challenging due to the inherent misalignment of feature spaces before and after unlearning. To address this, we introduce isolation verification distance (IVD) as a proxy metric that approximates statistical independence by analyzing the local feature structure. IVD leverages nearest-neighbor relationships within the feature space. The key insight is that if unlearning is effective, the forgotten sample should no longer share meaningful feature similarities with its previous closest neighbors in the retained dataset. Thus, IVD quantifies the unlearning at the feature level by measuring the change in nearest-neighbor distances: (1) for single entry data forgetting, IVD computes the feature distance between the forgotten data point and its nearest neighbors n in the retained dataset; (2) for batch data forgetting, IVD measures the distance of each forgotten data point to the class prototype of the retained data.

Algorithm 2 Single-Entry Forgetting Isolation Verification Distance (IVD)

Require: AdaptForget-induced deep learning model h_Θ, a query sample x_f, training dataset QO, number of nearest neighbors n

Ensure: IVD

1: **procedure** IVDAUDIT
2: Identify the n nearest neighbors of x_f in QO, denoted as $\{x_r^i\}_{i=1}^n$
3: Compute $\text{IVD}_{\text{Euclidean}}$, $\text{IVD}_{\text{Manhattan}}$, and $\text{IVD}_{\text{Score}}$ between $F_\theta(x_f)$ and $F_\theta(x_r^i)$
return $\text{IVD}_{\text{Euclidean}}$, $\text{IVD}_{\text{Manhattan}}$, $\text{IVD}_{\text{Score}}$

We use the IVD for single-entry data forgetting as an illustrative example, which is detailed as follows. Before the forgetting process, the feature similarity between the forgotten sample and its nearest neighbors is nearly identical. Effective unlearning requires that the forgotten sample is sufficiently isolated in the feature space. Thus, after the forgetting process, a larger feature distance between the forgotten sample and its n nearest neighbors in the feature space indicates more effective forgetting. This feature divergence serves as a key indicator of the unlearning process's success. As shown in Algorithm 2, the following metrics are employed: $\text{IVD}_{\text{Euclidean}}$, $\text{IVD}_{\text{Manhattan}}$, and a combined metric termed $\text{IVD}_{\text{Score}}$, which integrates these distances to provide a holistic measure of feature isolation:

$$\begin{aligned}
\text{IVD}_{\text{Euclidean}} &= \frac{1}{n}\sum_{i=1}^{n}\left\|F_\theta(x_f) - F_\theta(x_r^i)\right\|_2, \\
\text{IVD}_{\text{Manhattan}} &= \frac{1}{n}\sum_{i=1}^{n}\left\|F_\theta(x_f) - F_\theta(x_r^i)\right\|_1, \\
\text{IVD}_{\text{Score}} &= \alpha \cdot \text{IVD}_{\text{Euclidean}} + (1-\alpha)\cdot \text{IVD}_{\text{Manhattan}}.
\end{aligned} \tag{8}$$

Here, $F_\theta(x)$ represents the feature representation of the input x extracted by the model. The n nearest neighbors of x_f, denoted as x_r^i, are selected based on their proximity in feature space. The parameter α ensures a balanced assessment and is set to 0.5 in the experiments.

Novelty of IVD. While the formula combines ℓ_2 and ℓ_1 distances, IVD is novel as the first feature-level isolation audit metric for machine unlearning. Existing metrics (EMA p-value, accuracy, F1) operate exclusively at the output level and cannot detect feature-space vulnerabilities to attacks like feature inversion (Dosovitskiy, 2016) or attribute inference (Ganju, 2018). IVD fills this critical gap by quantifying feature-level isolation: whether forgotten samples remain entangled in the feature manifold with retained data, providing complementary auditing beyond output-level guarantees.

For output-level forgetting evaluation, we follow (Zhou, 2023), utilizing the EMA p-value, accuracy, and F1-score.

EMA p-value Calculation. The Ensembled Membership Auditing (EMA) p-value (Huang, 2021; Zhou, 2023) is a privacy auditing metric that quantifies the success of membership inference attacks. The calculation involves two steps: (1) Threshold inference: Using a calibration dataset (disjoint from the training set), we train a calibration model and compute three membership inference metrics: correctness (whether the model predicts correctly), confidence (the model's output logit for the true label), and entropy (the prediction entropy): for both training and test samples. For each metric, we select the threshold that maximizes $(TPR(t)+TNR(t))/2$, where TPR is the true positive rate and TNR is the true negative rate. (2) Membership testing: For each sample in the query dataset, we determine its membership status if at least one of the three metrics exceeds its corresponding threshold. We then perform a two-tailed Student's t-test between the membership prediction vector and an all-ones control vector. The resulting p-value serves as the audit metric: a low p-value (e.g., < 0.05) indicates that the attacker cannot distinguish forgotten samples from non-members, signifying successful forgetting. Importantly, each method's p-value is an independent privacy audit score (lower is better), not a comparative statistic between methods.

4. Experiments

To verify the robustness of AdaptForget in adaptive feature unlearning for trustworthy healthcare data privacy, we thoroughly examine the performance on batch data forgetting and single-entry data forgetting benchmark datasets against state-of-the-art methods with extensive analyses.

4.1. Setup

We evaluate AdaptForget on seven widely recognized healthcare datasets, covering both medical imaging and clinical health records to ensure broad applicability, as shown in Figure 2. The medical imaging datasets include PathMNIST, RetinaMNIST, DermaMNIST, and OCTMNIST from MedMNIST (Yang, 2023), spanning pathology, retinal fundus, dermatology, and OCT scans. The clinical health record datasets consist of Autism Spectrum Disorder (ASD) (Thabtah, 2017), MIMIC Death Dataset (MDD) (Johnson, 2016), and a Diabetes dataset (Smith, 1988). For the seven datasets, we randomly sampled partial data from the training dataset with percentages k from {0.25, 0.5, 0.75} for AdaptForget. We prepare query datasets with different sizes N from {1, 10, 100, 500, 1000, 2000}. A query dataset that completely overlaps with the training dataset is labeled as QO, while a query dataset that is completely disjoint with the training dataset is labeled QNO. In particular, we emphasize the single-entry data forgetting setting by focusing on QF where $N = 1$. See Appendix H for detailed dataset descriptions and split configurations.

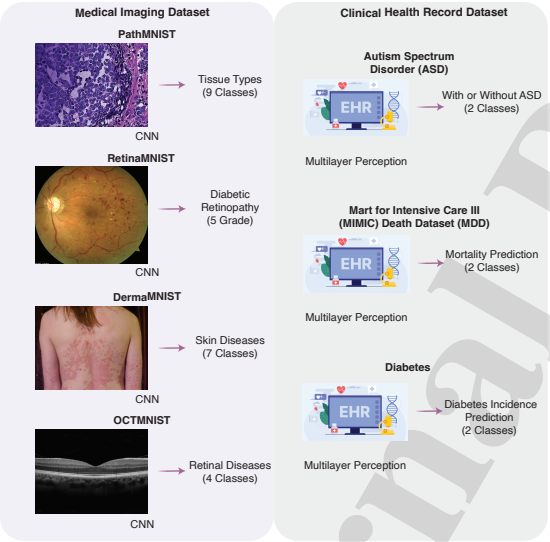


Figure 2: Illustration of seven datasets and deep learning (DL) models used to thoroughly evaluate both single-entry and batch data forgetting capabilities of the AdaptForget framework. Each dataset has unique characteristics that present distinct patient privacy in data forgetting.

We compare AdaptForget with a diverse set of state-of-the-art machine unlearning methods across both general and healthcare-specific domains. These include INS (Cha, 2024), which removes individual data points by adjusting model gradients; FisherForgetting (Golatkar, 2020), which selectively modifies model parameters based on Fisher information; SISA (Bourtoule, 2021), a shard-based approach that enables efficient unlearning by retraining only affected model partitions; and CF-K/EU-K (Goel, 2022), which employ kernel-based and ensemble strategies for approximating unlearning effects. Additionally, we evaluate Blindspot (Tarun, 2023), which relies on influence functions to estimate data removal effects; SSD-Tuning (Foster, 2024), a method that accelerates unlearning by selectively refining sub-networks; and Amnesiac unlearning (Graves, 2021), which applies data perturbations to erase memorized information. For healthcare-specific comparisons, we benchmark AdaptForget against AFS (Zhou, 2023), a teacher-student framework designed for medical applications.

We deploy two separate models for both medical image datasets and tabular datasets: a large DL model (teacher model) and a small DL model (student model). The teacher model refers to the original pre-trained model, while the student model is the new model generated by AdaptForget. For the medical image datasets, we employed convolutional neural networks (O'Shea, 2015). For the tabular datasets, we utilized multilayer perceptrons. The training process starts with a random initialization of the student model. The student model's parameters are updated through gradient descent, with the losses being weighted by specific balancing parameters. The learning rate for the randomly initialized part was set to 1e-5, and the Adam optimizer was employed. For the critical task of forgetting a single data point (QF=1), we highlight the importance of training a new student model for each instance to ensure fair and accurate verification of single-entry data forgetting. This process was repeated 100 times to thoroughly validate the effectiveness of AdaptForget in achieving precise data erasure.

4.2. Results

4.2.1. Single-entry Data Forgetting

AdaptForget achieves SOTA in single-entry forgetting across both clinical tabular datasets and medical imaging datasets (Figure 3). In the clinical datasets, such as *ASD* and *MDD*, AdaptForget achieves remarkable feature-level forgetting as evidenced by the IVD metrics. For instance, in the ASD dataset, AdaptForget achieves an $\text{IVD}_{\text{Euclidean}}$ of 8.941 and an $\text{IVD}_{\text{Manhattan}}$ of 53.246, significantly higher than FisherForgetting's 3.425 and 19.696, respectively. This trend is consistent across the Diabetes dataset, where AdaptForget achieves an $\text{IVD}_{\text{Score}}$ of 5.031, outperforming the next-best baseline, AFS, which only reaches 3.317. These results demonstrate AdaptForget's ability to ensure minimal residual impact of the forgotten data on model features, making it highly effective for privacy preservation in tabular datasets. In medical imaging datasets, AdaptForget further sets itself apart with its exceptional feature-level unlearning. For instance, in *RetinaMNIST*, AdaptForget achieves the highest $\text{IVD}_{\text{Euclidean}}$ (45.887) and $\text{IVD}_{\text{Manhattan}}$ (1022.039), significantly exceeding the values achieved by other methods, such as FisherForgetting (37.605 and 828.366, respectively). Similarly, in the *OCTMNIST* dataset, AdaptForget attains an $\text{IVD}_{\text{Score}}$ of 278.404, far surpassing SSD-Tuning's 96.8095 and FisherForgetting's 142.3105. This highlights AdaptForget's capability to completely remove the influence of forgotten samples not only from the output-level predictions but also from the deeper intrinsic feature representations. Additionally, AdaptForget achieves this feature-level forgetting while maintaining high accuracy and F1-score across datasets,

Table 2: Batch forgetting results for QF100 and QF1000.

(a) QF100

Method	ASD	Diabetes	MDD	OCTMNIST	RetinaMNIST	DermaMNIST	PathMNIST
EMA p-value							
FisherForgetting	2.76e-02	1.22e-04	1.77e-03	1.09e-01	5.86e-12	2.02e-03	1.49e-02
SISA	1.06e-01	3.52e-06	5.65e-02	1.35e-03	1.90e-10	1.21e-07	1.93e-04
CF-K	2.40e-01	1.46e-05	1.68e-02	1.07e-01	6.74e-04	7.47e-04	2.83e-02
EU-K	2.40e-01	2.16e-07	7.90e-03	2.86e-03	1.48e-05	2.33e-04	1.47e-02
Blindspot	5.48e-02	1.05e-01	5.02e-01	1.36e-01	2.98e-06	3.74e-03	5.23e-02
SSD-Tuning	1.22e-01	1.05e-01	1.59e-01	4.96e-02	2.32e-07	1.88e-03	2.75e-02
Amnesiac	3.88e-01	3.90e-04	1.23e-02	2.91e-02	4.77e-06	5.84e-04	1.92e-06
AFS (k=0.25)	4.16e-01	5.23e-02	1.49e-02	6.84e-04	3.61e-04	2.50e-03	4.22e-05
AFS (k=0.5)	6.67e-02	1.05e-07	6.77e-04	3.70e-03	1.18e-02	8.00e-03	1.55e-04
AFS (k=0.75)	5.45e-02	4.30e-03	5.29e-02	1.25e-03	3.60e-05	3.15e-02	4.30e-05
AdaptForget (k=0.25)	**3.50e-06**	**3.01e-09**	**1.62e-04**	**5.81e-06**	**3.92e-16**	**5.24e-22**	**1.04e-10**
AdaptForget (k=0.5)	6.09e-06	5.82e-08	2.86e-03	**1.25e-10**	9.38e-14	1.60e-18	1.05e-09
AdaptForget (k=0.75)	6.73e-06	2.16e-07	1.62e-04	1.72e-08	6.29e-10	2.51e-08	1.73e-08
IVD-Score							
FisherForgetting	0.58	0.31	0.41	223.65	274.28	234.30	199.82
SISA	16.18	5.26	6.05	148.86	147.52	200.96	136.53
CF-K	4.36	0.69	0.76	219.99	312.39	240.71	180.48
EU-K	4.36	0.69	0.76	201.56	279.16	221.42	177.09
Blindspot	0.60	0.28	0.44	219.50	272.95	232.20	190.10
SSD-Tuning	0.32	0.28	0.44	186.96	264.54	215.03	189.84
Amnesiac	0.69	0.31	0.44	192.34	254.21	176.22	129.92
AFS (k=0.25)	3.56	0.72	0.53	370.68	364.67	327.05	193.86
AFS (k=0.5)	2.30	0.75	0.68	340.40	178.58	306.87	198.03
AFS (k=0.75)	3.30	0.95	0.77	336.59	258.21	22.09	229.53
AdaptForget (k=0.25)	**50.82**	**6.67**	11.91	**1065.81**	564.70	**1039.04**	728.56
AdaptForget (k=0.5)	39.65	5.33	**31.20**	631.51	**848.47**	951.49	662.18
AdaptForget (k=0.75)	43.41	5.74	16.79	903.14	815.88	456.44	**728.56**
Accuracy							
FisherForgetting	93.75	72.58	70.73	70.89	45.83	71.44	77.13
SISA	94.33	72.41	70.73	64.90	49.17	72.38	81.31
CF-K	95.33	72.91	71.93	71.20	55.00	72.68	76.47
EU-K	95.33	72.79	71.60	69.37	51.38	71.83	77.96
Blindspot	94.05	73.00	70.82	71.28	53.13	70.37	74.54
SSD-Tuning	91.66	72.09	70.83	67.77	51.91	71.06	78.29
Amnesiac	93.15	72.69	69.91	65.29	53.38	70.92	51.91
AFS (k=0.25)	94.00	69.22	63.10	66.87	59.17	71.68	78.88
AFS (k=0.5)	93.00	70.52	65.31	70.45	59.17	73.45	81.28
AFS (k=0.75)	94.33	70.90	67.00	70.91	56.67	74.81	80.24
AdaptForget (k=0.25)	95.00	73.01	70.47	69.17	57.08	72.05	80.68
AdaptForget (k=0.5)	**96.00**	73.32	71.67	72.25	61.88	75.07	83.44
AdaptForget (k=0.75)	95.83	**73.32**	**72.00**	**72.51**	**62.17**	**75.40**	**85.72**
F1-Score							
FisherForgetting	0.9314	0.7260	0.7062	0.6944	0.4355	0.7188	0.7724
SISA	0.9605	0.7226	0.7068	0.5660	0.2854	0.3415	0.7569
CF-K	0.9422	0.7286	0.7156	0.6782	0.3793	0.4773	0.7300
EU-K	0.9422	0.7272	0.7158	0.6522	0.3226	0.4898	0.7286
Blindspot	0.9348	0.7300	0.7033	0.7038	0.5269	0.7156	0.7513
SSD-Tuning	0.9178	0.7190	0.7027	0.6576	0.5094	0.7132	0.7868
Amnesiac	0.9439	0.7260	0.6978	0.5895	0.4872	0.6853	0.4663
AFS (k=0.25)	0.9595	0.7124	0.5940	0.5975	0.3378	0.3446	0.7333
AFS (k=0.5)	0.9496	0.7218	0.6141	0.6691	0.3736	0.4236	0.7641
AFS (k=0.75)	0.9620	0.7286	0.6586	0.6730	0.3838	0.4585	0.7977
AdaptForget (k=0.25)	0.9667	0.7322	0.7031	0.6540	0.5250	0.7018	0.8000
AdaptForget (k=0.5)	**0.9733**	0.7287	0.7159	0.6991	0.6007	0.7260	0.8267
AdaptForget (k=0.75)	0.9720	**0.7354**	**0.7191**	**0.7038**	**0.6060**	**0.7412**	**0.8526**

(b) QF1000

Method	ASD	Diabetes	MDD	OCTMNIST	RetinaMNIST	DermaMNIST	PathMNIST
EMA p-value							
FisherForgetting	8.17e-02	5.33e-46	2.89e-18	3.13e-20	2.13e-96	2.82e-24	4.33e-32
SISA	6.26e-02	3.76e-14	2.18e-04	1.02e-04	7.97e-76	5.65e-46	2.11e-43
CF-K	5.41e-01	9.40e-25	1.68e-05	2.30e-03	2.65e-22	7.69e-16	7.69e-16
EU-K	5.41e-01	5.52e-35	7.01e-19	6.73e-04	2.28e-33	2.53e-26	4.08e-15
Blindspot	5.26e-02	7.65e-30	2.47e-30	7.81e-16	3.64e-44	1.57e-24	2.13e-28
SSD-Tuning	2.16e-03	1.32e-22	3.33e-01	1.25e-14	6.02e-53	1.34e-15	1.29e-27
Amnesiac	1.81e-01	3.13e-50	2.91e-04	9.67e-20	9.50e-69	1.88e-39	1.43e-40
AFS (k=0.25)	2.30e-03	5.52e-35	1.19e-10	1.00e-06	1.03e-16	9.96e-22	4.53e-36
AFS (k=0.5)	3.17e-02	4.65e-27	2.68e-11	2.75e-26	1.92e-16	1.33e-28	3.49e-30
AFS (k=0.75)	3.36e-01	1.55e-37	2.01e-11	1.62e-25	9.75e-60	3.03e-24	1.92e-04
AdaptForget (k=0.25)	**3.00e-04**	**2.10e-53**	**4.74e-47**	1.70e-52	**9.66e-130**	3.17e-49	1.05e-53
AdaptForget (k=0.5)	5.30e-03	5.28e-46	1.55e-37	**1.33e-66**	1.87e-56	5.31e-54	8.48e-36
AdaptForget (k=0.75)	2.30e-03	4.35e-41	6.63e-34	6.02e-62	1.76e-75	**1.46e-58**	**2.84e-56**
IVD-Score							
FisherForgetting	0.56	7.61	0.39	247.54	274.28	279.67	182.96
SISA	6.04	19.45	7.55	146.07	96.74	126.66	102.45
CF-K	1.81	7.35	0.75	243.85	387.06	253.38	155.64
EU-K	1.81	6.17	0.75	231.59	279.16	240.03	135.73
Blindspot	0.57	7.39	0.42	241.60	259.67	254.47	160.72
SSD-Tuning	0.66	7.28	0.42	241.62	282.55	272.25	163.53
Amnesiac	0.52	5.17	0.41	169.78	282.56	268.83	99.02
AFS (k=0.25)	1.16	11.81	0.22	332.20	282.37	285.75	179.90
AFS (k=0.5)	1.43	13.32	0.12	325.93	288.08	301.98	205.18
AFS (k=0.75)	1.18	12.07	0.16	320.24	288.43	295.65	190.82
AdaptForget (k=0.25)	6.98	40.65	8.12	**1605.68**	1592.85	**1515.54**	1109.35
AdaptForget (k=0.5)	7.36	**55.63**	**9.21**	1177.15	971.25	920.12	719.88
AdaptForget (k=0.75)	**10.36**	36.66	7.08	1073.61	**1591.52**	1057.75	**1293.43**
Accuracy							
FisherForgetting	95.09	73.83	70.79	70.86	45.83	71.39	77.72
SISA	94.00	73.18	70.35	65.00	47.78	72.38	81.90
CF-K	94.50	73.68	71.80	71.63	52.93	75.08	83.14
EU-K	94.50	72.99	71.55	71.23	51.39	69.39	78.56
Blindspot	95.98	73.92	70.39	70.27	54.43	68.36	75.95
SSD-Tuning	85.27	73.93	70.52	72.16	52.08	71.01	78.66
Amnesiac	95.53	72.61	71.48	63.31	52.60	72.28	78.41
AFS (k=0.25)	93.67	72.62	67.70	67.50	59.20	71.98	77.22
AFS (k=0.5)	91.33	73.54	70.00	72.00	59.52	74.21	79.51
AFS (k=0.75)	93.70	74.18	69.50	72.13	60.56	74.91	82.61
AdaptForget (k=0.25)	95.00	74.18	69.33	69.00	57.50	75.22	78.97
AdaptForget (k=0.5)	95.33	74.13	70.30	71.65	60.41	75.00	83.55
AdaptForget (k=0.75)	**96.00**	**74.56**	**71.87**	**72.30**	**61.67**	**75.30**	**83.55**
F1-Score							
FisherForgetting	0.9500	0.7381	0.7067	0.6957	0.4355	0.7184	0.7776
SISA	0.9200	0.7311	0.7033	0.5630	0.2344	0.3505	0.7707
CF-K	0.9300	0.7359	0.7177	0.6850	0.3661	0.4762	0.7779
EU-K	0.9300	0.7287	0.7152	0.6648	0.3227	0.4582	0.7541
Blindspot	0.9630	0.7389	0.6985	0.6940	0.5457	0.7055	0.7651
SSD-Tuning	0.8570	0.7390	0.7003	0.7111	0.5149	0.7145	0.7896
Amnesiac	0.9580	0.7227	0.7117	0.5857	0.4972	0.7164	0.7651
AFS (k=0.25)	0.9500	0.7251	0.6813	0.6089	0.3684	0.3681	0.7030
AFS (k=0.5)	0.9420	0.7311	0.6920	0.6850	0.4007	0.4328	0.7359
AFS (k=0.75)	0.9550	0.7326	0.6965	0.6867	0.3972	0.4476	0.7764
AdaptForget (k=0.25)	0.9689	0.7418	0.6926	0.6655	0.5304	0.7323	0.7816
AdaptForget (k=0.5)	0.9688	0.7359	0.6999	0.6882	0.6004	0.7263	**0.8262**
AdaptForget (k=0.75)	**0.9748**	**0.7436**	**0.7184**	**0.7111**	**0.6103**	**0.7363**	0.8247

ensuring the utility of the model is not compromised. For example, in *PathMNIST*, AdaptForget achieves an accuracy of 0.794 and an F1-score of 0.992, both significantly higher than baselines.

To demonstrate the completeness of AdaptForget in single-entry data forgetting, we compare the results of AdaptForget with those of a retrained model (i.e., a model trained without the forgotten data points) on single-entry data forgetting. The retrained model serves as a gold standard, representing a model trained from scratch without the forgotten sample. We calculate the Kullback–Leibler (KL) divergence between the output probability distributions of the two models for the forgotten data points (see Appendix D for a full discussion). After sequentially forgetting 100 samples, the KL divergences obtained by AdaptForget on PathMNIST, RetinaMNIST, DermaMNIST, OCTMNIST, ASD, MDD, and Diabetes datasets were 0.35, 0.26, 2.38e-06, 0.81, 2.51e-08, 3.43e-08, and 1.92e-08, respectively. These values are very close to those of the retrained model, indicating that AdaptForget effectively minimizes the impact of the forgotten data points, achieving reliable forgetting performance.

We validate AdaptForget's effectiveness in forgetting rare disease patient data using the MIMIC-III Death Dataset. We focus on acute renal failure with tubular necrosis (ICD-9 5845), a severe condition requiring privacy protection due to potential risks to patients and genetically related individuals. AdaptForget was applied to forget data related to these patients, including diagnostic information and fatal disease indicators (e.g., hyperkalemia ICD-9 2767, congestive heart failure ICD-9 4280). Experimental results show that AdaptForget effectively isolates forgotten data in the feature space, as confirmed by the IVD metric. The Euclidean and Manhattan distances ($\text{IVD}_{\text{Euclidean}} = 2.16$, $\text{IVD}_{\text{Manhattan}} = 11.09$) indicate significant feature separation. Furthermore, AdaptForget achieves near-identical performance to a retrained model (accuracy 0.716, F1-score 0.709), ensuring effective and complete feature-level forgetting. This case study serves as an initial proof-of-concept for single-entry forgetting within a specific rare-condition subgroup.

4.2.2. *Batch Data Forgetting*

AdaptForget demonstrates significant advantages over multiple baseline methods across medical imaging datasets. Through comprehensive evaluations across critical metrics: p-value, normalized $\text{IVD}_{\text{Score}}$, accuracy, and F1-score, AdaptForget shows

its ability to achieve robust forgetting while maintaining high model performance (Table 2). Specifically, the results reveal three key points: (1) In privacy protection, AdaptForget consistently achieves low p-values for membership attack success rates, reflecting its robust ability to prevent data leakage. For example, in the *PathMNIST* dataset, AdaptForget achieves a p-value of 4.71×10^{-27} for forgetting 100 query samples (QF100), significantly outperforming competitors like AFS (2.87×10^{-6}) and SISA (2.15×10^{-3}). (2) In feature isolation, AdaptForget ensures effective separation of forgotten data from retained data in the high-dimensional feature space, as reflected by consistently high normalized $\text{IVD}_{\text{Score}}$ across datasets. For instance, in the *OCTMNIST* dataset, AdaptForget achieves the highest normalized $\text{IVD}_{\text{Score}}$ for both QF100 and QF1000, outperforming AFS and FisherForgetting, which struggle to achieve similar levels of feature-level forgetting. (3) In performance preservation, AdaptForget maintains competitive accuracy and F1-score, even as the scale of forgotten data increases. For instance, in the *PathMNIST* dataset, AdaptForget achieves an accuracy of 0.8580 and an F1-score of 0.9762 for QF1000, showing minimal performance degradation compared to the original model. This trend is consistent across datasets like *DermaMNIST*, where AdaptForget achieves an F1-score of 0.8036 for QF1000, outperforming baseline methods while retaining strong model utility.

AdaptForget also demonstrates SOTA performance on clinical health record datasets (Table 2), achieving effective data forgetting and ensuring robust feature-level unlearning. These datasets: ASD (identifying autistic traits in toddlers), MDD (predicting patient mortality), and Diabetes (predicting diabetes incidence), represent diverse clinical applications where privacy preservation is paramount. For instance, in the ASD dataset, AdaptForget achieves a p-value of 5.73×10^{-6} for QF1000 with $k = 0.25$, while maintaining an accuracy of 0.9533 and an F1-score of 0.9677. Additionally, the $\text{IVD}_{\text{Manhattan}}$ distance indicates high feature isolation (1330.175), confirming the model's ability to erase sensitive traits at the feature level. For the Diabetes dataset, AdaptForget achieves a p-value of 5.46×10^{-44} for QF1000 while maintaining high accuracy (0.7289) and F1-score (0.7269), alongside strong feature-level performance with an $\text{IVD}_{\text{Manhattan}}$ distance of 324.9384, ensuring that the erased data is effectively disentangled from retained features. These results highlight AdaptForget's ability to balance output-level accuracy and feature-level forgetting.

4.3. Analysis and Discussion

Ablation on teacher-student learning paradigm. When compared to independent teacher and independent student models, the performance of AdaptForget highlights the unique advantages of its teacher-student framework (Figure 3a). In PathMNIST, AdaptForget achieves consistently lower p-values in the ensembled membership attack for both QF100 and QF1000 settings. For instance, when $k = 0.75$, AdaptForget achieves a p-value of 1.73×10^{-8} for QF100 and 2.84×10^{-56} for QF1000, compared to 9.70×10^{-5} and 1.75×10^{-20} for the independent teacher model, and 5.18×10^{-4} and 5.66×10^{-31} for the independent student model. In $k = 0.25$, AdaptForget achieves an $\text{IVD}_{\text{Score}}$ of 918.9525, significantly higher than the independent teacher (159.395) and student (296.45).

Table 3: Single-entry Data Forgetting Results across All Metrics.

Method	ASD	Diabetes	MDD	OCTMNIST	RetinaMNIST	DermaMNIST	PathMNIST
IVD-Score							
FisherForgetting	11.5605	2.6410	4.9960	142.3105	432.9855	291.9685	232.9660
Amnesiac	10.0915	2.5935	2.6395	75.5750	368.2200	235.9695	144.3415
Blindspot	10.9225	2.5435	4.7245	131.0895	370.9770	284.3600	225.2970
SSD-Tuning	6.9860	1.4660	2.0675	96.8095	199.2870	210.0700	214.6270
AFS	29.5935	3.3170	3.6400	274.4710	472.9640	442.3160	330.5375
AdaptForget	**31.0935**	**5.0310**	**5.5425**	**278.4040**	**533.9630**	**537.4195**	**547.6480**
IVD-Manhattan							
FisherForgetting	19.696	4.342	8.370	271.039	828.366	558.085	444.391
Amnesiac	17.116	4.101	4.242	142.372	702.371	451.439	274.171
Blindspot	18.586	4.164	7.893	249.413	708.885	543.510	430.219
SSD-Tuning	11.804	2.385	3.432	183.217	379.970	401.673	409.419
AFS	51.535	5.555	6.125	525.038	906.665	847.519	634.170
AdaptForget	**53.246**	**7.749**	**9.097**	**532.017**	**1022.039**	**1025.675**	**1049.314**
IVD-Euclidean							
FisherForgetting	3.425	0.940	1.622	13.582	37.605	25.852	21.541
Amnesiac	3.067	1.086	1.037	8.778	34.069	20.500	14.512
Blindspot	3.259	0.923	1.556	12.766	33.069	25.210	20.375
SSD-Tuning	2.168	0.547	0.703	10.402	18.604	18.467	19.835
AFS	7.652	1.079	1.155	23.904	39.263	37.113	26.905
AdaptForget	**8.941**	**2.313**	**1.988**	**24.791**	**45.887**	**49.164**	**45.982**
Accuracy							
FisherForgetting	0.9046	0.6307	0.6066	0.6483	0.3535	0.5558	0.7375
Amnesiac	0.8509	0.7254	0.6960	0.7272	0.4479	0.7097	0.7435
Blindspot	0.9453	0.7255	0.6813	0.7034	0.4583	0.7140	0.7755
SSD-Tuning	0.9529	0.7238	0.6941	0.7379	0.4704	0.7214	0.6116
AFS	0.9367	0.7289	0.7123	0.7509	0.5549	0.7246	0.7292
AdaptForget	**0.9536**	**0.7299**	**0.7315**	**0.7541**	**0.5840**	**0.7250**	**0.7940**
F1-Score							
FisherForgetting	0.897	0.5874	0.5508	0.6018	0.3056	0.5463	0.7370
Amnesiac	0.8399	0.7248	0.6940	0.6943	0.4382	0.7140	0.7490
Blindspot	0.9452	0.7252	0.6763	0.6782	0.4443	0.7185	0.7750
SSD-Tuning	0.9522	0.7232	0.6919	0.7017	0.3990	0.6588	0.5960
AFS	0.9554	0.7278	0.7209	0.8864	0.2018	0.6026	0.9370
AdaptForget	**0.9672**	**0.7283**	**0.7352**	**0.9091**	**0.5016**	**0.8252**	**0.9920**

Analysis on scalability and efficiency. AdaptForget's scalability on batch data forgetting is evident in its performance across datasets with varying retaining data scales. The parameter k is varied among {0.25, 0.5, 0.75} to assess the impact of different levels of data retention in the student models, while the size of QF is also adjusted to understand its effect on the unlearning process. As shown in Table 2, in the *OCTMNIST* dataset, AdaptForget achieves a p-value of 3.56×10^{-15} for QF100 and maintains an accuracy of 0.7288 when $k = 0.25$. As k increases to 0.75, the p-value increases to 1.95×10^{-13} for QF100, while the accuracy improves to 0.7613. Similarly, increasing QF size from 100 to 1000 samples while keeping $k = 0.5$, AdaptForget maintains a low p-value of 9.03×10^{-19}, with an F1-score of 0.8949. In Figure 3b, AdaptForget-generated models demonstrate lower GPU memory usage during inference compared to the original models. In tabular datasets like Diabetes and ASD, it drops significantly from 0.48 MB to 0.08 MB, which is crucial for deployment in resource-constrained environments. Figure 3g further illustrates the comprehensive performance of AdaptForget across five key dimensions: ability to forget, accuracy, efficiency, dataset size, and model size, demonstrating a balanced solution.

Ablation on OOD-FOD loss. Figures 3 (c-f) collectively highlight the significant advantages of the newly-proposed OOD-Driven Feature-Output Disentanglement (OOD-FOD) loss. Figures 3c shows that using OOD-FOD loss consistently outperforms the naive method that randomizes the features of forgotten data in both forgetting effectiveness and model per-

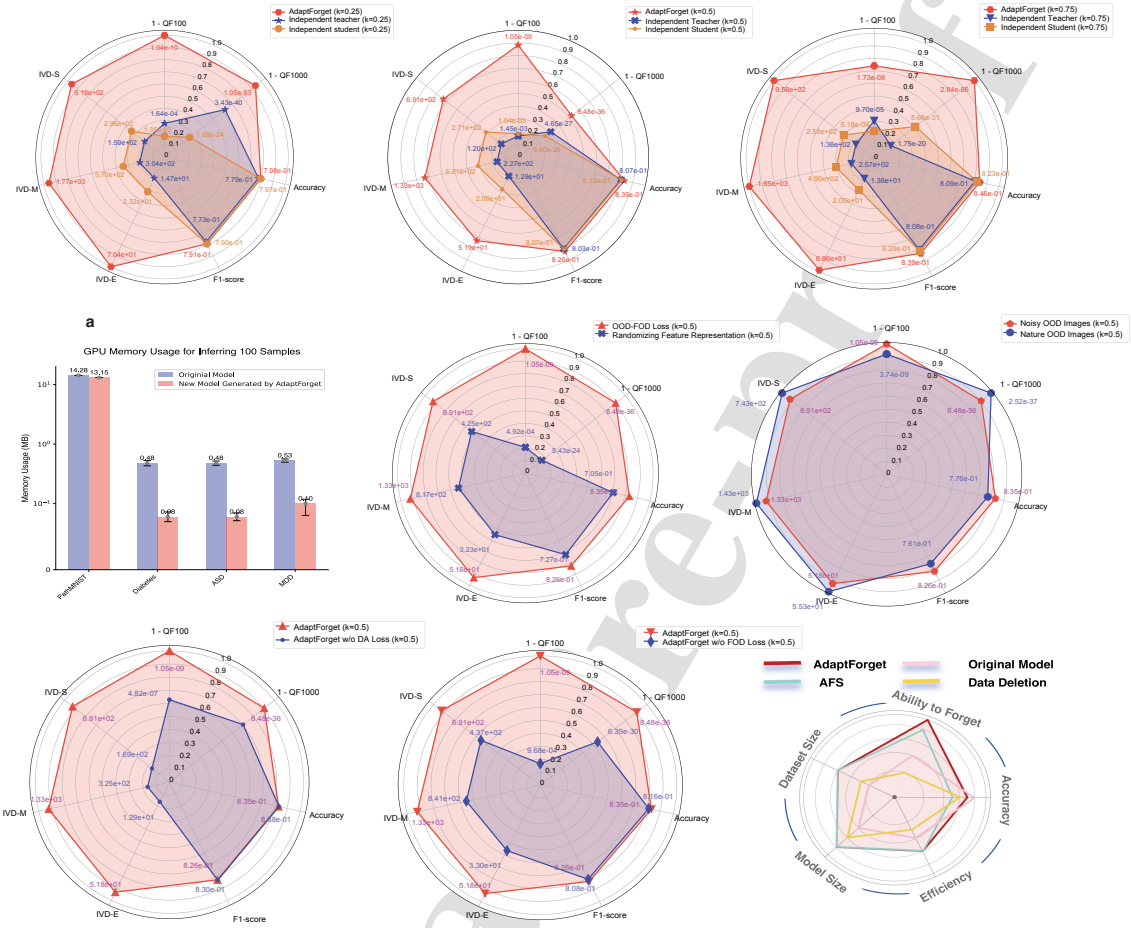


Figure 3: Illustration of various potential ablation methods and strategies. Metrics include QF100, QF1000, representing the p-value of ensembled membership attack success rates for forgetting 100 or 1,000 samples. IVD-M, IVD-E, and IVD-S denote Manhattan distance, Euclidean distance, and a combined score of these distances for IVD. All metrics except accuracy and F1-score were normalized. All labeled values are real experimental results. **a** Compares the performance of the teacher-student model on the PathMNIST dataset with $k = 0.25, 0.5, 0.75$. **b** Analyzes GPU memory usage during inference of 100 samples. **c** Compares domain adaptation to randomizing feature representation. **d** Examines the effect of using noisy samples as out-of-distribution (OOD) data compared to natural images as OOD data. **e** Ablation study showing the impact of omitting domain adaptation and OOD data on performance. **f** Ablation study showing the effect of removing model-guided prediction randomization loss. **g** Summarizes overall performance across five key dimensions.

formance, as demonstrated on the PathMNIST dataset with a retaining sample ratio of $k = 0.5$. Specifically, OOD-FOD loss achieves significantly lower QF100 p-value (1.05×10^{-9} vs. 4.92×10^{-4}) and QF1000 p-value (8.48×10^{-36} vs. 5.43×10^{-24}), indicating its ability to ensure more thorough unlearning. Since OOD-FOD loss has two parts, excluding the first domain adversarial (DA) loss significantly compromises forgetting effectiveness, as shown in Figure 3e. The QF100 and QF1000 p-values increase dramatically (from 1.05×10^{-9} to 4.82×10^{-7} and 8.48×10^{-36} to 1.44×10^{-32}, respectively), indicating weaker unlearning. Furthermore, the Euclidean and Mahalanobis distances drop significantly (from 51.8 to 12.9 and from 1330.18 to 324.94, respectively), highlighting a loss in feature-level isolation of forgotten data. Figure 3f further investigates the effect of removing the loss of the second feature-output separation (FOD). Without this loss, the QF100 p-value increases from 1.05×10^{-9} to 9.68×10^{-4}, and the QF1000 p-value rises from 8.48×10^{-36} to 6.39×10^{-30}, showing a substantial reduction in forgetting efficiency.

Analysis on the tradeoff on OOD data setting. In PathMNIST experiments, we compare different OOD data settings, contrasting the use of natural images as OOD data with our approach of adding noise to create OOD samples. The results, summarized in Figure 3d, demonstrate that using natural images as OOD

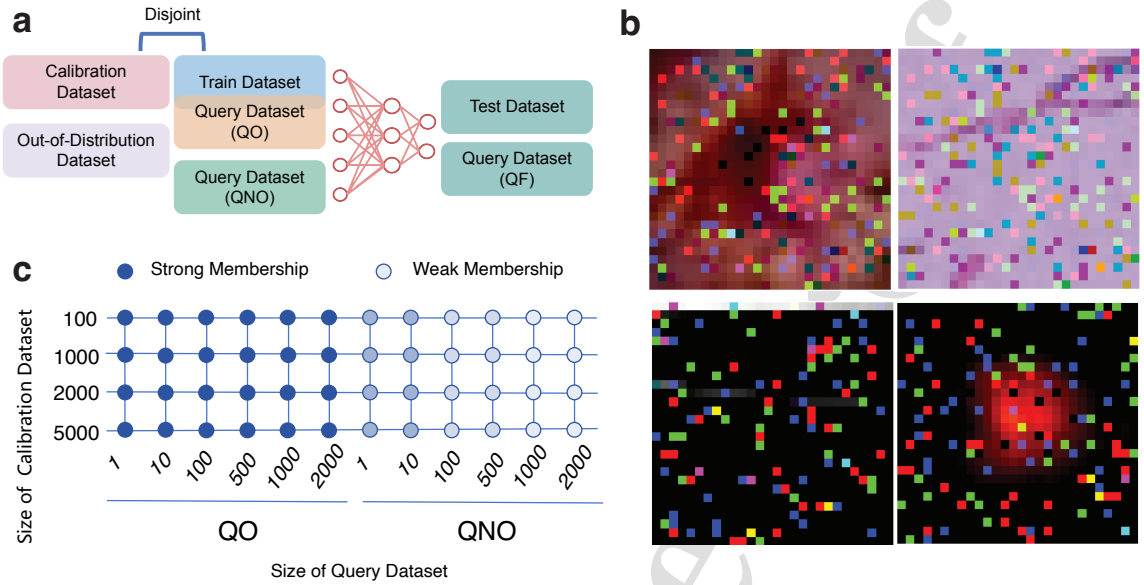


Figure 4: **a** illustrates the disjoint relationships between the calibration dataset, training dataset, and different query datasets (QO, QNO, QF). **b** Examples of the Out-of-Distribution (OOD) data generated by Gaussian noise. **c** shows P-values obtained from auditing models trained with different query datasets (QO and QNO), highlighting the model's strong and weak membership across varying query dataset sizes.

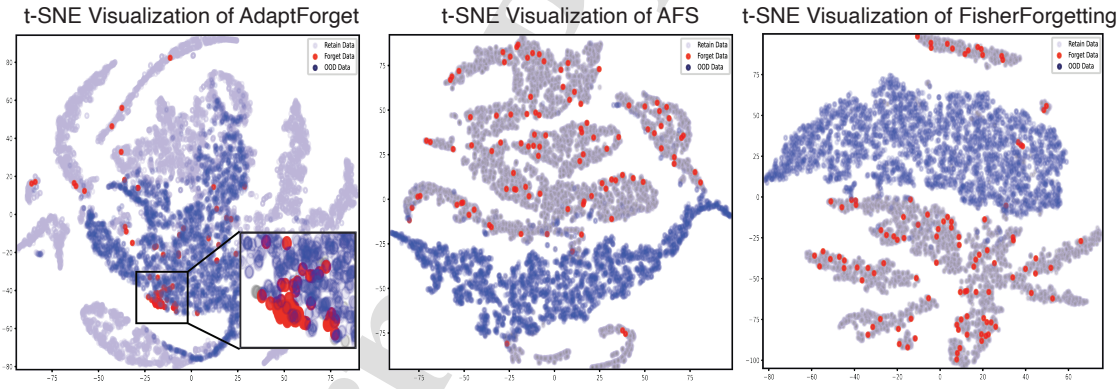


Figure 5: The t-SNE visualizations of feature representations show that AdaptForget effectively aligns the forgotten samples and OOD samples (see Appendix B for a full comparison of methods).

significantly leads to degraded accuracy and F1-score, suggesting that natural images cause the medical data distribution to deviate too far from its original feature space, impairing model outputs. In contrast, our method, which generates OOD data by corrupting training data with Gaussian noise (Figure 4b), represents a balanced approach. This strategy enables the model to achieve effective feature-level forgetting while maintaining strong model performance metrics, including accuracy and F1-score. The corrupted images ensure that OOD samples remain within a reasonable distance from the original medical data distribution, effectively promoting a controlled forgetting process.

To further validate the robustness of noise-based OOD selection, we conduct ablation studies across eight OOD types: gaussian, uniform, salt_pepper, poisson, stain_weak, stain_medium, stain_strong, and deepmind. We position Gaussian noise as a practical, verifiable baseline with three key properties: (1) no semantic prior: noise distributions lack task-relevant structure, reducing the risk of inadvertently preserving class-discriminative features; (2) parameter controllability: noise intensity can be systematically varied for reproducible ablation studies; (3) cross-dataset generalizability: unlike domain-specific style transfer, noise-based OOD does not require

Table 4: Hyperparameter sensitivity analysis. Each cell shows test accuracy / AUC and EMA p-value (blue). Underlined values indicate the best result per sub-table.

(a) Fixed $\lambda_1 = 1.0$ std(Acc) = 2.31%

λ_2 (RL)	λ_3 (KP)					
	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	**81.74** / 0.817 1.1e-05	**84.46** / 0.840 1.4e-06	**83.38** / 0.828 2.2e-05	**84.18** / 0.838 3.3e-07	**83.10** / 0.825 1.1e-05	**84.50** / 0.841 1.6e-08
0.2	**78.00** / 0.770 4.8e-10	**79.02** / 0.785 4.8e-10	**81.36** / 0.809 3.3e-14	**79.08** / 0.784 3.4e-12	**80.82** / 0.801 3.3e-14	**79.62** / 0.789 4.8e-10
0.4	**75.30** / 0.743 4.9e-08	**78.94** / 0.787 4.9e-08	**80.02** / 0.794 3.4e-12	**79.50** / 0.788 4.8e-10	**80.14** / 0.798 4.8e-10	**80.92** / 0.799 2.2e-08
0.6	**76.42** / 0.756 4.9e-08	**79.36** / 0.787 2.2e-09	**79.88** / 0.791 2.2e-08	**80.48** / 0.801 4.8e-10	**79.40** / 0.787 1.3e-10	**80.76** / 0.802 4.8e-10
0.8	**76.20** / 0.759 3.4e-12	**79.46** / 0.791 3.4e-12	**79.70** / 0.789 3.1e-11	**79.10** / 0.786 3.4e-12	**81.32** / 0.806 1.4e-17	**79.66** / 0.789 3.1e-11
1.0	**73.42** / 0.734 2.2e-09	**78.90** / 0.783 1.4e-17	**80.44** / 0.799 3.1e-11	**79.22** / 0.785 1.3e-09	**78.86** / 0.782 4.8e-10	**79.56** / 0.786 3.1e-11

(b) Fixed $\lambda_2 = 0.0$ std(Acc) = 1.30%

λ_1 (DA)	λ_3 (KP)					
	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	**79.52** / 0.790 3.3e-07	**82.66** / 0.823 6.9e-07	**84.14** / 0.834 1.6e-07	**83.94** / 0.824 1.4e-06	**82.64** / 0.822 2.2e-05	**82.62** / 0.821 1.4e-06
0.2	**81.14** / 0.805 2.9e-06	**84.20** / 0.837 4.4e-05	**82.86** / 0.822 1.6e-07	**82.80** / 0.823 3.3e-07	**83.02** / 0.823 1.6e-07	**83.56** / 0.831 5.8e-06
0.4	**79.58** / 0.795 6.9e-07	**84.18** / 0.838 5.8e-06	**83.36** / 0.828 2.9e-06	**82.58** / 0.821 3.3e-07	**82.52** / 0.821 1.1e-05	**83.02** / 0.826 1.6e-08
0.6	**82.50** / 0.816 2.2e-05	**84.40** / 0.837 6.9e-07	**83.84** / 0.836 1.6e-07	**83.28** / 0.828 6.9e-07	**82.96** / 0.822 2.9e-06	**83.36** / 0.828 1.4e-06
0.8	**78.98** / 0.787 3.3e-07	**82.80** / 0.822 1.1e-05	**83.40** / 0.830 3.3e-07	**83.08** / 0.827 3.3e-07	**83.32** / 0.825 3.3e-07	**82.70** / 0.829 5.8e-06
1.0	**81.74** / 0.817 1.1e-05	**84.46** / 0.840 1.4e-06	**83.38** / 0.828 2.2e-05	**84.18** / 0.838 3.3e-07	**83.10** / 0.825 1.1e-05	**84.50** / 0.841 1.6e-08

(c) Fixed $\lambda_3 = 0.8$ std(Acc) = 1.16%

λ_1 (DA)	λ_2 (RL)					
	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	**82.64** / 0.822 2.2e-05	**82.00** / 0.814 1.3e-09	**80.86** / 0.805 1.3e-09	**80.54** / 0.798 1.3e-09	**82.04** / 0.813 4.8e-10	**80.80** / 0.800 3.4e-12
0.2	**83.02** / 0.823 1.6e-07	**80.56** / 0.800 2.2e-08	**80.62** / 0.798 1.3e-09	**79.94** / 0.795 4.8e-10	**79.88** / 0.789 3.4e-12	**80.02** / 0.796 3.1e-11
0.4	**82.52** / 0.821 1.1e-05	**80.90** / 0.803 3.1e-11	**81.48** / 0.811 3.1e-11	**80.42** / 0.794 3.1e-11	**81.08** / 0.804 3.4e-12	**79.10** / 0.783 1.4e-17
0.6	**82.96** / 0.822 2.9e-06	**80.88** / 0.802 3.2e-07	**80.72** / 0.800 4.8e-10	**80.40** / 0.796 3.1e-11	**79.80** / 0.792 1.3e-09	**80.54** / 0.800 1.3e-09
0.8	**83.32** / 0.825 3.3e-07	**80.92** / 0.805 2.2e-08	**80.50** / 0.796 3.1e-11	**80.06** / 0.795 4.8e-10	**78.94** / 0.785 2.2e-09	**80.26** / 0.792 1.3e-09
1.0	**83.10** / 0.825 1.1e-05	**80.82** / 0.801 3.3e-14	**80.14** / 0.798 4.8e-10	**79.40** / 0.787 1.3e-10	**81.32** / 0.806 1.4e-17	**78.86** / 0.782 4.8e-10

dataset-specific generative models. Across all noise types, forget-set features migrate away from the original cluster and disperse toward the OOD distribution, demonstrating that noise mapping truly alters the feature manifold rather than merely masking it.

Analysis on forget-set accuracy. To directly quantify the degree of forgetting, we evaluate the post-unlearning model's performance on the forget set itself using Forget-set Accuracy (Forget_Acc), defined as the percentage of correct predictions on forgotten samples (lower values indicate more thorough forgetting). We conduct validation experiments on the Camelyon17-WILDS dataset, a clinical pathology benchmark with 224×224 resolution images that present realistic challenges including higher input dimensionality, complex textural features, and multi-center heterogeneity typical of clinical deployments. Table 5 shows the results. Baseline methods retain high Forget_Acc (AFS: 99.33%, SISA: 97.33%, CF-K: 92.67%), indicating strong memory residue. In contrast, AdaptForget achieves Forget_Acc = 63.0% with Test_Acc = 90.34% (MIA p-value: 1.0×10^{-12}), demonstrating effective forgetting while maintaining high diagnostic performance. The discrepancy between Forget_Acc (63.0%) and Test_Acc (90.34%) is a strong indicator of selective erasure: the model has specifically lost the "fingerprint" of the forgotten samples without global performance degradation.

Table 5: Performance comparison on Camelyon17.

Method	Forget Acc ↓	Test Acc ↑	IVD-S ↑	MIA ratio ↓	MIA p-value
AdaptForget	63.00	90.34	61.65	**63.00**	**9.96e-13**
AFS	99.33	84.71	59.50	99.33	0.5457
Amnesiac	79.00	47.57	18.21	88.67	0.0024
Blindspot	82.33	67.35	7.57	94.67	0.0424
CF-K	92.67	90.35	17.51	93.00	0.0069
EU-K	92.67	89.84	17.43	92.67	0.0067
Fisher	**46.67**	46.98	**1.60e5**	100.00	1.0000
INS	97.33	**94.18**	39.15	97.33	0.1663
SISA	97.33	90.21	22.25	97.33	0.1479
SSD	55.00	57.09	10.98	88.67	0.3334

Analysis on the feature disentanglement. The t-SNE visualizations in Figure 5d illustrate the capability of AdaptForget to achieve robust feature-level forgetting while preserving the alignment of forgotten and OOD samples. In the visualiza-

tion of AdaptForget (left panel), the forgotten samples (red) are closely clustered with OOD samples (blue), indicating successful disentanglement from the retained data (purple). This alignment demonstrates that AdaptForget effectively directs the unlearning process within the high-dimensional feature space by leveraging the interplay of retaining data, forgotten data, and OOD data as inputs. In contrast, the visualizations of AFS and FisherForgetting show that nearly 99% of the forgotten samples remain entangled with the retained samples, demonstrating significantly weaker feature disentanglement. This indicates that these methods fail to achieve effective feature isolation and comprehensive unlearning compared to AdaptForget. For a comprehensive comparison with additional methods, see Appendix B.

Analysis on various query size. The analysis in Figure 4c demonstrates the impact of query dataset size on AdaptForget's performance under different scenarios. QO, QNO, and QF are distinct query datasets designed to evaluate various aspects of the forgetting mechanism (Figure 4a). Specifically, QF overlaps with the original training dataset, allowing for the evaluation of ensembled membership attack. In contrast, QNO consists of samples that do not belong to the original training set, serving as a proxy for test data. QO includes samples overlapping with the retaining data, enabling the evaluation of whether the forgetting process adversely affects the retaining data. In the QNO scenario, as the query dataset size increases from 1 to 2000 samples (Figure 4c), AdaptForget exhibits an enhanced ability to differentiate between training and non-training datasets, as evidenced by the progression from strong membership (dark blue) to weak membership (light blue). This indicates effective forgetting of the targeted data while maintaining the integrity of unrelated datasets. In contrast, under the QO scenario, where the query data overlaps with the retaining data, the model's performance on retaining data remains unaffected (Figure 4c). The consistent forgetting effect demonstrates that AdaptForget avoids catastrophic forgetting and preserves the retaining data's predictive performance, regardless of the query dataset size.

Ablation on hyperparameters. To assess the impact of the hyperparameters λ_1, λ_2, and λ_3 in the loss function, we conducted an ablation study on the PathMNIST dataset, analyzing their effects on accuracy, F1-score, and p-value. As shown in Table 4, AdaptForget exhibits relatively stable performance around the default setting (std $= 2.31$ percentage points across 36 tested configurations of λ_2, λ_3), with bounded degradation only at design extremes. Specifically, the minimum accuracy of 73.42% occurs at the single boundary configuration ($\lambda_2 = 1.0$, $\lambda_3 = 0.0$) where the knowledge preservation loss is completely disabled; all other 35 configurations exceed 78%, with most achieving 80–84%. The F1-score is mostly within 0.78–0.84, and the p-value remains low across all tested configurations, suggesting robust forgetting under different hyperparameter choices.

To further investigate cross-modal generalizability, we conduct systematic hyperparameter sensitivity analysis across imaging (Camelyon17) and tabular (ASD) modalities. For λ_1 (knowledge distillation weight), imaging data shows moderate sensitivity with test accuracy ranging 69.25–85.38% (CV=8.71%), while tabular data demonstrates high stability with accuracy ranging 91.67–93.33% (CV=0.79%). For λ_2 (domain adversarial weight), imaging data exhibits test accuracy range of 69.25–87.88% (CV=10.54%), whereas tabular maintains 91.67–95.00% (CV=1.59%). These results indicate that a shared hyperparameter set ($\lambda_1 = 1.0$, $\lambda_2 = 1.0$) serves as a practical default across both modalities, with imaging data showing moderate sensitivity and tabular demonstrating substantially higher stability (Figure 6).

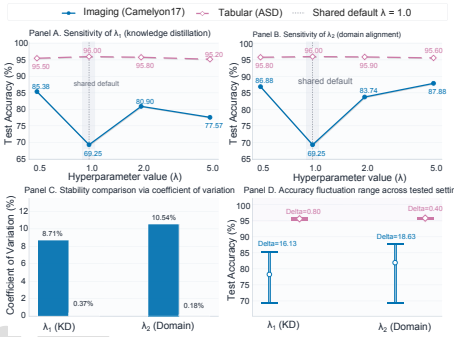


Figure 6: Cross-modal hyperparameter sensitivity analysis. Panels A-B show imaging modality (Camelyon17) sensitivity to λ_1 and λ_2. Panels C-D show tabular modality (ASD) stability across the same hyperparameter ranges.

Comparative sensitivity analysis with baseline methods. To further validate AdaptForget's hyperparameter robustness, we conduct a comparative sensitivity analysis against 9 baseline unlearning methods (AFS, Amnesiac, Blindspot, FisherForgetting, EU-K, CF-K, SISA, SSD-Tuning, INS), testing each across the same hyperparameter ranges: learning rates from 10^{-4} to 2×10^{-3} and varying training steps. As illustrated in Figure 7, AdaptForget showed 0% training collapse across all 34 configurations (Forget_Acc: 45–75%, Test_Acc predominantly $> 80\%$), forming a tight cluster within the stable region (green shaded area). In contrast, baseline methods exhibited varying instability: FisherForgetting has 100% collapse rate, AFS showed 24% collapse with high variance (Test_Acc: $65.6 \pm 12.5\%$), and SSD-Tuning had 20% collapse. This relative stability stems from our representation-level approach, which operates at the feature level through domain adaptation, providing smoother optimization compared to weight-space methods that directly perturb model parameters.

Validation on feature attacks. After applying AdaptForget, existing feature attack methods, such as feature inversion attacks (Dosovitskiy, 2016; Mahendran, 2015), and attribute inference attacks (Ganju, 2018), show a significant reduction in success rates, demonstrating the robustness of the forgetting mechanism. Following feature inversion attacks (Dosovitskiy, 2016), which attempt to reconstruct original images from the model's internal representations, we evaluate their performance on the PathMNIST dataset. Before forgetting, the inversion attack reconstructed images with a visual similarity score (SSIM) of 0.78 on average. However, after forgetting the specific data samples, the SSIM drops to 0.14. Following attribute inference

Figure 7: Privacy-utility trade-off comparison. AdaptForget (green star, $n = 34$) occupies a stable region (green shaded area) with all configurations clustering tightly. Baseline methods (colored dots, $n = 25$ each) show scattered distributions, with collapsed configurations marked as $\times$. Large dots represent method-wise means.

attacks (Ganju, 2018), which aim to infer sensitive attributes from model outputs, we observe a dramatic decrease in the success rate on the tabular dataset MDD. Before forgetting, the success rate of inferring attributes is 44.8%, but after forgetting the targeted samples using AdaptForget, the success rate drops to 8.1%.

Attribute-level forgetting validation. To further validate clinical applicability, we conduct an attribute-level forgetting experiment on Camelyon17, where the `node` label (hospital/center ID) serves as a proxy for sensitive institutional attributes. We designate samples from `node=1` as the forget set, with remaining nodes as the retain set. We evaluate three metrics jointly: (1) SSIM Inversion Risk ($\text{SSIM}_{\text{forget}}$, lower is better) measuring reconstruction quality under white-box inversion attacks, (2) Attribute Identifiability ($\text{attr}_{\text{forget}}$, lower is better) measuring accuracy of a post-hoc attribute classifier trained on unlearned model features (random baseline: 20% for 5-class balanced classification), and (3) Task Performance (test_{acc}, higher is better) measuring diagnostic accuracy on the full test set.

Table 6: Attribute-level forgetting results showing privacy-utility trade-offs across all compared methods.

Method	$\text{SSIM}_{\text{forget}}$	$\text{attr}_{\text{forget}}$	test_{acc}
AdaptForget	0.2625	13.50%	90.86%
Fisher	**0.127**	1.60%	25.19%
SSD	0.356	**0.00%**	50.00%
Amnesiac	0.279	30.85%	47.10%
Blindspot	0.257	41.75%	49.43%
SISA	0.277	54.20%	86.76%
INS	0.263	55.25%	**92.08%**
CF-K	0.280	57.40%	90.28%
EU-K	0.280	59.20%	90.21%

As shown in Table 6, AdaptForget is the only method that reduces attribute identifiability below the random baseline (13.50% < 20%) while maintaining high diagnostic accuracy (90.86%). Methods with comparable task performance (INS: 92.08%, CF-K: 90.28%, EU-K: 90.21%) fail to erase the attribute (55-59% identifiability). Aggressive forgetting methods (Fisher, SSD) achieve low attribute identifiability but their task accuracy collapses to 25-50%, representing model failure rather than selective forgetting.

Cross-domain unlearning validation. To validate AdaptForget's applicability to federated-like scenarios, we conducted a domain-level unlearning experiment on Camelyon17-WILDS multi-center dataset. We simulate a scenario where three medical centers (Centers 0, 3, 4) collaboratively train a pathology classification model, after which Center 4 withdraws due to privacy compliance and requests data removal. We train the teacher model on Centers 0, 3, 4, then forget all samples from Center 4 while retaining Centers 0 and 3. Table 7 shows AdaptForget achieves a favorable balance between forgetting efficacy (Forget Acc: 57.38%) and task utility (Test F1: 0.887), outperforming baseline methods that either fail to forget effectively (SISA: 86.58% Forget Acc) or sacrifice task performance (FISHER: 0.1359 Test F1).

Table 7: Cross-domain unlearning results comparison.

Method	Forget Acc (%)	Test F1
AdaptForget	57.38	**0.887**
FISHER	**39.72**	0.1359
SSD	50.94	0.6667
AMNESIAC	52.5	0.7135
INS	62.15	0.8577
BLINDSPOT	63.24	0.2594
EU_K	64.73	0.8573
CF_K	66.79	0.8696
SISA	86.58	0.8291

This cross-domain experiment demonstrates AdaptForget's applicability to federated learning scenarios. Unlike conventional federated unlearning methods (e.g., FedEraser) that require synchronous multi-party collaboration and substantial communication overhead, AdaptForget operates as a server-side post-federated unlearning solution. This enables the central server to efficiently remove a withdrawing party's data contribution without re-engaging all participating hospitals: a critical advantage in medical contexts where coordinating retraining cycles is often logistically infeasible. Furthermore, by targeting feature-level representations rather than output re-calibration, AdaptForget provides robust defense against sophisticated privacy threats like feature inversion attacks, offering a practical zero-interaction unlearning mechanism for federated medical imaging systems.

Analysis on metric correlation. Figure 3f demonstrates the correlation between the feature-level audit metric (IVD) and the output-level audit metric (ensembled membership attack, EMA). Specifically, the scatter plots show positive correlations between $\text{IVD}_{\text{Score}}$ and EMA ($R^2 = 0.429$), $\text{IVD}_{\text{Euclidean}}$ and EMA ($R^2 = 0.424$), and $\text{IVD}_{\text{Manhattan}}$ and EMA ($R^2 = 0.420$) across all datasets. These results indicate that while the IVD metrics and EMA are not identical, they exhibit a consistent alignment. This correlation underscores the ability of IVD to reflect meaningful changes in feature-level representations dur-

ing unlearning. More importantly, IVD offers complementary insights that EMA alone cannot capture, as it evaluates the forgetting process from a granular feature-level perspective.

Sequential unlearning stability. To evaluate long-horizon behavior under repeated deletion requests, we conduct a sequential unlearning protocol where the same deployed model is updated cumulatively over 100 consecutive forget requests. Across these 100 steps, AdaptForget maintains stable utility without catastrophic collapse in the tested setting. Retain accuracy remains essentially unchanged (94.28% $\rightarrow$ 94.95%), while test accuracy exhibits only a modest decrease (92.84% $\rightarrow$ 91.69%, -1.15 percentage points). Meanwhile, IVD-S increases from 34.42 to 66.83 (+94%), indicating progressively stronger feature-space isolation of forgotten samples. Since this increase is not accompanied by a collapse in retain or test accuracy, it is consistent with selective forgetting being strengthened rather than uncontrolled feature drift. For a matched-step comparison, we evaluate SISA at 30 sequential requests (computational constraints prevent extending to 100 steps). SISA achieves higher retain accuracy (97.44%), reflecting the benefit of shard retraining, but AdaptForget attains stronger feature isolation (IVD-S: 55.96 vs 30.26) and competitive test accuracy with a substantially lighter update mechanism. We therefore view AdaptForget as offering a favorable efficiency-stability trade-off for practical settings with frequent deletion requests. Figure 8 visualizes the trajectory across 100 steps.

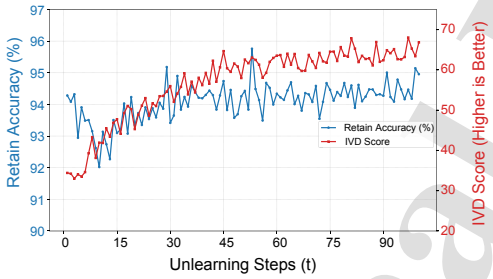


Figure 8: Longitudinal analysis of 100 sequential unlearning steps. Retain Accuracy (blue) remains stable, while IVD-S (red) increases, indicating strengthening representation-level forgetting without utility collapse.

To further validate IVD's utility as a feature-space audit metric, we examine its correlation with feature-based attacks (Feature Inversion measured by SSIM). Based on cross-method comparison (10 methods), we observe a strong negative correlation: Pearson $r = -0.92$, $R^2 = 0.85$, $p = 1.65 \times 10^{-4}$. As shown in Figure 9, IVD exhibits significantly stronger correlation with feature reconstruction risk (SSIM, $R^2 = 0.85$) compared to its correlation with output-level membership inference (EMA, $R^2 = 0.42$). This is consistent with IVD's design: it measures feature-space isolation, which directly relates to feature-level privacy leakage. The weak IVD-EMA correlation reflects that feature-level and output-level privacy risks operate at different abstraction layers; a method can achieve strong output-level privacy (low EMA) while still leaking feature-level information (low IVD), or vice versa.

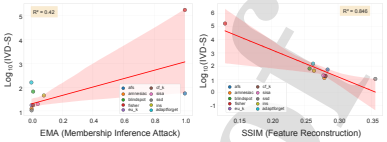


Figure 9: IVD-S correlation with feature-level attack (SSIM). IVD-S shows strong negative correlation with SSIM ($R^2 = 0.85$, $p < 0.001$), indicating that higher feature isolation corresponds to lower reconstruction quality. Each point represents a different unlearning method.

Subgroup analysis on retained data. To evaluate whether unlearning causes disproportionate impacts on specific subgroups within the retained set, we conduct per-class evaluation on Camelyon17. Table 8 and Figure 10 show the results.

Table 8: Subgroup performance on retained set before and after unlearning.

Retained Subgroup	Before	After	Delta
Non-Tumor sensitivity	99.90%	78.19%	-21.71 pp
Non-Tumor F1-score	0.998	0.873	-0.125
Tumor sensitivity	99.81%	99.12%	-0.69 pp
Tumor F1-score	0.999	0.902	-0.097
Overall accuracy	99.85%	88.95%	-10.90 pp

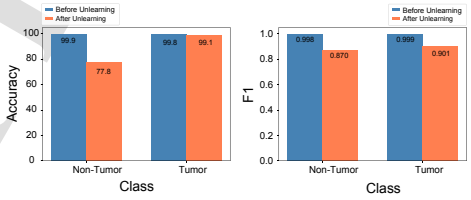


Figure 10: Class-wise performance comparison on retained set. Left: Per-class accuracy (sensitivity). Right: Per-class F1-score. The unlearning process causes disproportionate degradation on the Non-Tumor subgroup while preserving Tumor detection capability.

Performance degradation is not uniform across classes. Non-Tumor sensitivity drops by 21.71 percentage points, while Tumor sensitivity remains largely preserved with only 0.69 percentage point decrease. Confusion matrix analysis reveals that only 9/1028 Tumor samples (0.87%) are misclassified as Non-Tumor (false negatives), while 212/972 Non-Tumor samples (21.8%) are misclassified as Tumor (false positives). This indicates the degradation primarily manifests as reduced specificity rather than reduced sensitivity.

To verify whether the local manifold structure of retained samples remains intact, we conducted k-NN analysis. As shown in Figure 11, the k-NN accuracy on retained samples drops by only 2.47%-2.84% across different k values (5, 10, 20), indicating that the local manifold structure remains largely stable. This demonstrates that while the classifier's decision boundaries have shifted, the neighborhood relationships of retained samples in the embedding space have not been severely disrupted.

Visual interpretability analysis. To validate attention pattern changes, we apply Grad-CAM Selvaraju (2017) to visualize

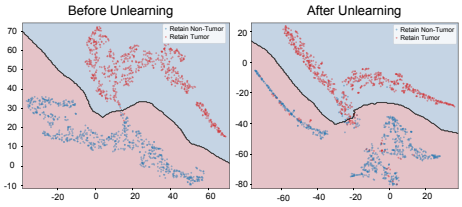


Figure 11: k-NN manifold structure analysis on retained samples. The visualization shows decision boundary comparison before and after unlearning, demonstrating that local neighborhood relationships remain largely intact despite boundary shifts.

model focus before and after unlearning. Quantitative analysis shows: (1) IoU@top20%=0.00, indicating complete relocation of peak attention regions; (2) cosine similarity=0.37±0.24, confirming substantial shift in spatial attention distribution; (3) confidence drop=0.26±0.03 on forget samples. Figure 12 demonstrates that post-unlearning attention disperses from discriminative features (e.g., tumor regions) to background areas, mechanistically explaining the performance degradation on forget classes.

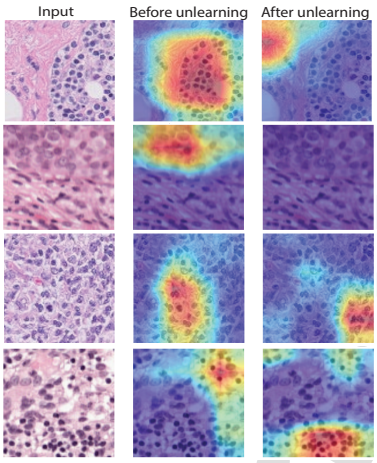


Figure 12: Grad-CAM visualization comparing attention patterns before and after unlearning. The heatmaps show that the model's attention shifts from discriminative regions to diffuse background areas after unlearning.

Output probability distribution analysis. To quantify the uncertainty of model predictions on forgotten data, we analyze the prediction entropy distribution. Results show that after unlearning, mean entropy increases from 0.035 (5% of maximum) to 0.693 (99.9% of maximum entropy), with output probabilities approaching uniform distribution ([50.6%, 49.4%]) and KL divergence of 0.0006. This demonstrates that the model achieves maximum uncertainty on forgotten data, consistent with treating them as out-of-distribution samples (Figure 13).

Extension to nature image forgetting. The results in Table 9 demonstrate that AdaptForget consistently outperforms existing unlearning methods across both CIFAR-10 and MNIST. The model achieves the highest IVD score, indicating superior feature disentanglement, and maintains high classification accuracy and F1-score, ensuring minimal utility degradation. This highlights its potential for broader application in natural image forgetting, where high-dimensional feature entanglement presents significant challenges. The use of OOD-driven feature-output disentanglement makes AdaptForget adaptable to diverse data distributions, enabling its extension to settings of real-world scene datasets.

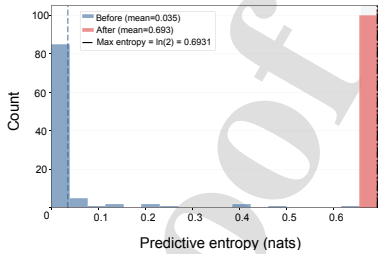


Figure 13: Prediction entropy distribution before and after unlearning. After unlearning, the entropy distribution concentrates near maximum entropy $\ln(2) = 0.693$, indicating maximum uncertainty.

Validation of OOD selection strategy. To validate the empirical OOD selection strategy proposed in Section 3.3, we conducted multi-configuration analysis on the Camelyon17-WILDS dataset. Table 10 demonstrates that while OOD intensity acts as a lever to balance privacy and utility, unlearning performance remains stable across multiple noise types. Specifically, Gaussian noise with low intensity achieves Forget Acc = 63.0% and OOD AUROC = 0.88, demonstrating strong forgetting effectiveness. Increasing the intensity to medium further reduces Forget Acc to 55.0%, but at the cost of reduced Test Acc (84.9%). Uniform noise with low intensity provides a balanced alternative with Forget Acc = 73.0% and Test Acc = 91.7%. This confirms that the proposed strategy provides a practical starting point for empirically selecting OOD parameters, with dataset-specific tuning remaining necessary.

5. Conclusion

In this work, we introduced AdaptForget, a domain-adaptive feature-level machine unlearning framework designed to ensure complete and verifiable forgetting while maintaining model utility. Unlike traditional output-level unlearning approaches, AdaptForget employs a structured feature disentanglement strategy, leveraging out-of-distribution (OOD) data to guide the unlearning process in a controlled manner. This enables effective removal of forgotten data from both feature and output spaces while preventing feature collapse. To formalize the unlearning process, we established a feature-level unlearning generalization bound, providing theoretical insights into the trade-off between forgetting effectiveness and model generalization. We also proposed the OOD-driven feature-output disentanglement loss, which enforces feature alignment between forgot-

Table 9: Performance comparison of different unlearning methods on CIFAR-10 and MNIST datasets. The best results are in **bold**.

Method (k=0.5)	**CIFAR-10**				**MNIST**			
	P-value	Acc	F1	IVD	P-value	Acc	F1	IVD
AdaptForget	**6.88e-07**	**0.7788**	**0.7783**	**329.945**	**3.31e-07**	**0.9860**	**0.9860**	**347.962**
AFS	1.14e-05	0.7754	0.7739	308.802	6.90e-03	0.9778	0.9774	338.149
Amnesiac	1.61e-04	0.7358	0.7320	97.0341	1	0.9802	0.9805	118.851
Blindspot	2.01e-03	0.7697	0.7710	224.144	1	0.9820	0.9822	202.553
CF-K	1.56e-01	0.7712	0.7725	130.449	1	0.9848	0.9847	191.210
EU-K	8.44e-05	0.7728	0.7729	65.387	1	0.9840	0.9840	42.294
FisherForgetting	5.78e-04	0.7279	0.7253	214.441	3.18e-01	0.9814	0.9815	225.315
SISA	1.37e-05	0.7278	0.7277	185.298	3.18e-01	0.9746	0.9745	140.691
SSD-Tuning	6.90e-03	0.7703	0.7655	241.922	1	0.9854	0.9854	203.112
INS	2.86e-06	0.5930	0.5921	204.414	1.08e-03	0.7672	0.7502	210.906

Table 10: Multi-configuration OOD analysis on Camelyon17.

OOD Type	Intensity	Forget Acc ↓	Test Acc ↑	F1 ↑
Uniform Noise	Low	73.0	91.7	0.91
Gaussian Noise	Low	63.0	90.3	0.91
Gaussian Noise	Medium	55.0	84.9	0.86

ten and OOD data while ensuring prediction-level randomization to break decision boundary dependencies. Furthermore, to address the lack of rigorous verification of unlearning, we introduced the isolation verification distance (IVD), a novel feature-level audit metric that quantitatively measures the extent of forgetting in latent representations. Extensive experiments across various datasets demonstrate that AdaptForget outperforms state-of-the-art unlearning methods in both batch and single-entry data forgetting while preserving model utility. Future work includes extending AdaptForget to more complex real-world scenarios, such as continuous learning, federated learning, and systematic validation on broader sensitive patient subgroups, where efficient and adaptive unlearning remains a challenge.

Acknowledgements

This project is supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No REI/1/5234-01-01, REI/1/5414-01-01, REI/1/5289-01-01, REI/1/5404-01-01, REI/1/5992-01-01, URF/1/4663-01-01, Center of Excellence for Smart Health (KCSH), under award number 5932, and Center of Excellence on Generative AI, under award number 5940, the National Natural Science Foundation of China (No. 62372280), the Natural Science Foundation of Shandong Province (No. ZR2024MF139, ZR2022QG051, and ZR2023QF094), Demonstration Projects of Science and Technology for the People of Qingdao City (No. 23-2-8-smjk-2-nsh), the Special fund of Qilu Health and Health Leading Talents Training Project, the Qilu Young Scholars Program, the Shandong Provincial Higher Education Young Innovation Team Development Program (2025KJH105), and the 2025 Postgraduate Quality Improvement and Innovation Project of Shandong University of Traditional Chinese Medicine [No. YJSTZCX2024007], China.

Appendix A. Details on the Feature Entanglement

Feature-level disentanglement is a critical aspect of effective unlearning, as it ensures that forgotten data points are not only removed from the model's predictions but are also functionally and representationally isolated in the feature space. This section discusses two key observations from Figure A.1, which highlight the limitations of existing methods and the advantages of AdaptForget.

Limitations of Existing Methods in Feature-Level Forgetting. From Figure A.1, it is evident that all existing unlearning methods, except AdaptForget, fail to achieve comprehensive feature-level forgetting. Approximately 99% of the forgotten samples remain entangled with the retaining samples in the high-dimensional feature space. This persistent entanglement indicates that these methods are unable to effectively isolate the feature representations of forgotten data, leaving significant residual dependencies. Such entanglement undermines the privacy guarantees of these methods, as the forgotten data points continue to influence the model's decision-making indirectly. This observation confirms that current approaches are inadequate for achieving complete feature-level unlearning, thereby compromising the integrity of the forgetting process.

Feature-Level Forgetting Achieved by AdaptForget. In contrast, AdaptForget demonstrates a clear separation of forgotten data points from the retaining samples. In Figure A.1, the forgotten data points are predominantly entangled with the OOD samples, while only a small fraction remains entangled with the retaining data. This pattern highlights the success of AdaptForget in aligning forgotten samples with OOD data, effectively removing their representational impact from the training data. The minimal entanglement with retaining samples further validates the effectiveness of AdaptForget in isolating feature representations. By leveraging domain-invariant representation learning and feature-level isolation, AdaptForget achieves a robust and comprehensive unlearning process, setting a new standard in privacy-preserving machine unlearning. These observations highlight the critical advantages of AdaptForget over existing methods, as it addresses the persistent feature entanglement challenge, ensuring that forgotten data points are truly erased at both the prediction and feature levels.

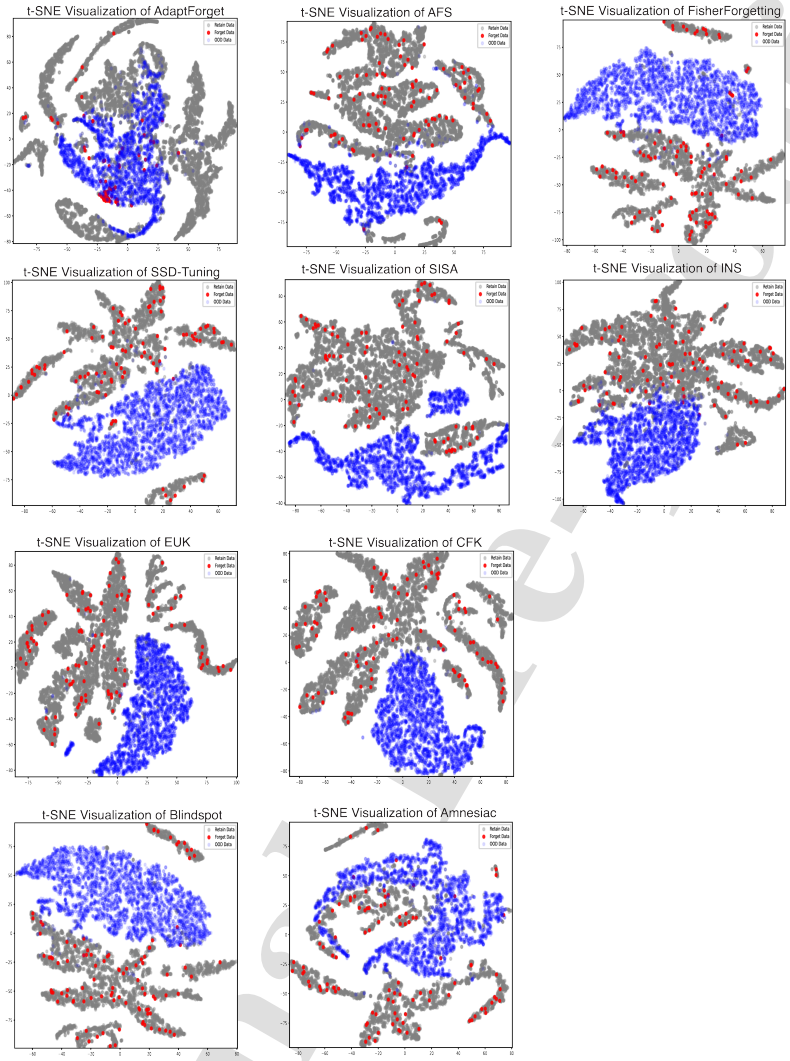


Figure A.1: t-SNE visualizations comparing feature representations of forgotten data, retaining data, and OOD data across various unlearning methods.

Appendix B. Full Discussion on Single-Entry Data Forgetting

The right to revoke personal healthcare data is a fundamental aspect of patient privacy. In dynamic healthcare environments, ensuring the real-time removal of individual patient records is paramount. Single-entry data forgetting addresses this need by enabling real-time deletion without waiting for a batch of data to accumulate. This capability is particularly critical in scenarios involving rare diseases, highly sensitive medical conditions, or histories of substance abuse, where timely compliance with data removal requests can significantly impact patient trust and privacy.

Single-entry data forgetting is uniquely tailored to erase individual patient records while maintaining the integrity of the remaining data. This approach safeguards vulnerable individuals from potential privacy violations, which could result in discrimination, stigmatization, or even personal harm. For instance, in dermatology-based diagnostic systems (Zhou, 2024), images of skin conditions are intrinsically tied to patient identities. If this information is not adequately erased, it risks being leaked, potentially causing irreversible harm to individuals' reputations and social standing.

As patients increasingly demand control over their personal information, healthcare organizations face mounting pressure to comply with deletion requests promptly and effectively. Failure to do so exposes these organizations to significant legal and eth-

ical liabilities, undermining public trust in AI-driven healthcare systems. Developing robust algorithms for single-entry data forgetting is vital for compliance and preserving the utility and fairness of healthcare models.

While single-entry data forgetting can be seen as a specific case of batch data forgetting, it introduces unique challenges that warrant independent consideration. DL models inherently entangle individual data points with the broader dataset in high-dimensional feature spaces. This entanglement complicates the process of erasing a single data point completely, especially when traditional machine unlearning methods focus primarily on adjusting output-level predictions. These methods often overlook the intrinsic feature representations, leaving the forgotten data's traceable fingerprints in the model. The detail challenges are as follows.

- **Complex Interdependencies:** DL models learn statistical patterns and interdependencies across the entire training dataset rather than treating data points in isolation (Serra, 2018; Van de Ven, 2020). The removal of a single data point can inadvertently disrupt these learned relationships, potentially altering the model's overall structure and performance.
- **Feature Space Clustering:** In the feature space, data points within the same class often form dense clusters (Cai, 2018; Kauffmann, 2024). Forgetting a single entry without disrupting the integrity of these clusters is challenging. Removing one point risks altering the model's representation of similar points, potentially degrading its predictive accuracy for related samples.
- **Dynamic Model Behavior:** DL models exhibit dynamic behavior during training, with individual data points fluctuating between being remembered and forgotten as the model continuously updates its internal states (Zhou, 2023; Huang, 2021). This dynamic nature complicates the verification process, making it difficult to ensure that a single data point has been fully erased without residual influence.

The AdaptForget framework directly addresses these challenges through domain-invariant representation learning and feature-level isolation. Unlike existing methods that focus solely on output-level predictions, AdaptForget effectively isolates forgotten data points in the feature space, pulling them closer to out-of-distribution (OOD) representations. This ensures that the forgotten data points are not only removed from the model's predictions but are also functionally and representationally distanced from the retained data. AdaptForget demonstrates that it is possible to achieve real-time unlearning of single data points while maintaining the model's performance on the retained data. By addressing the entanglement of data in high-dimensional feature spaces, it overcomes the limitations of traditional unlearning methods. This makes it particularly well-suited for healthcare applications, where the stakes of privacy violations are high, and timely compliance with data deletion requests is critical.

Appendix C. Detailed Discussion on the Limitations of Existing Metrics

Current machine unlearning tools are predominantly evaluated using metrics such as accuracy, F1-score, and ensembled membership attack (EMA), which assesses p-values of correctness, confidence, and entropy (Zhou, 2023). While these metrics effectively measure forgetting in the output space, they fall short in evaluating the feature-level unlearning. This creates a significant gap in ensuring comprehensive data protection, as improvements in output-level metrics do not necessarily equate to successful erasure of sensitive information in the feature representations.

Feature representations in deep learning models capture intricate patterns and characteristics of training data. Retention of such representations, even after unlearning, leaves models vulnerable to feature-space attacks like feature inversion (Dosovitskiy, 2016; Mahendran, 2015), attribute inference (Ganju, 2018), and gradient leakage (Zhu, 2019; Geiping, 2020). These attacks can reconstruct sensitive patient information from the internal states of the model, bypassing output-level defenses. Therefore, the absence of robust feature-space metrics undermines the effectiveness of existing unlearning methods, especially in privacy-sensitive domains such as healthcare.

EMA, a widely used audit mechanism, also exhibits notable limitations in auditing single-entry data forgetting and feature-level unlearning. While effective for batch data forgetting, EMA struggles to confirm the removal of individual data points. This limitation arises because the influence of a single data point on the model's overall behavior is often subtle, leading to minor changes in the model's outputs that EMA may fail to detect. Additionally, the inherent randomness and redundancy in deep neural networks obscure the contributions of single data points, making EMA less reliable in these scenarios. These shortcomings are particularly problematic in complex models, where such subtle changes can easily go unnoticed.

Feature-level auditing is another weak area for EMA. Although it monitors output-layer changes effectively, output-level forgetting does not guarantee the erasure of underlying feature representations. This results in "feature entanglement," where the features of forgotten data remain strongly associated with those of other data points. Such entanglement undermines the effectiveness of unlearning, leaving the model vulnerable to attacks like feature inversion and gradient leakage, which exploit the retained feature-level information. EMA's primary focus on output-layer metrics neglects these risks, providing an incomplete assessment of unlearning effectiveness.

To address these shortcomings, objective metrics for the feature space, such as Euclidean and Manhattan distances, should be incorporated. These metrics quantitatively measure the similarity between feature representations of the model before and after unlearning. For example, a large Euclidean distance between the latent representations of the forgotten data in the original model and the unlearned model indicates successful feature-level unlearning, whereas a minimal distance suggests residual retention of the data.

In addition, composite metrics like the IVD-Score, which

combines Euclidean and Manhattan distances, can offer a more nuanced evaluation of feature-space changes. By incorporating these metrics, unlearning methods can be evaluated comprehensively, ensuring both output predictions and internal representations no longer retain traces of the forgotten data. This approach addresses the limitations of EMA, enhancing the robustness of privacy protection mechanisms in machine unlearning.

Appendix D. Model Prediction Divergence Comparison

To assess the impact of the forgetting process on model predictions, we compare the outputs of the unlearning model and the retrained model for the same forgotten sample. The retrained model serves as a gold standard, representing a model trained from scratch without the forgotten sample. Since retraining a model without the forgotten data is computationally expensive, this comparison provides a benchmark to evaluate the success of the unlearning process. The smaller the prediction divergence for the forgotten sample between the unlearning model and the retrained model, the more successful the forgetting process. This comparison ensures that the unlearning process not only alters the feature representation but also significantly impacts the model's predictions for the forgotten sample.

To measure the distance between the probability distributions predicted by the unlearning model and the retrained model for the forgotten data point, the KL divergence is used to quantify this distance:

$$\begin{aligned}\mathrm{IVD}_{\mathrm{KL}} = \sum_{j=1}^{K} & C_{\text{student}}(F_{\text{student}}(x_{\text{forgotten}}))_j \\ & \times \log \frac{C_{\text{student}}(F_{\text{student}}(x_{\text{forgotten}}))_j}{C_{\text{retrain}}(F_{\text{retrain}}(x_{\text{forgotten}}))_j}.\end{aligned} \tag{D.1}$$

Here, $C_{\text{student}}(F_{\text{student}}(x))$ and $C_{\text{retrain}}(F_{\text{retrain}}(x))$ denote the output probability distributions of the unlearned model and the retrained model for input x, respectively, and K is the number of classes. A lower $\mathrm{IVD}_{\mathrm{KL}}$ indicates more effective forgetting of a specific data point, confirming the effectiveness of the forgetting process.

Appendix E. Theoretical Preliminaries

Definition 6 (Empirical $\mathcal{H}\Delta\mathcal{H}$-Discrepancy)**.** *Let $\mathcal{H}$ be a class of binary (or real-valued) hypotheses, and let $\widehat{P}$ and $\widehat{Q}$ be two* empirical distributions *(samples) of size m_P and m_Q from distributions P and Q over the same domain $\mathcal{X}$. Define*

$$\begin{aligned}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\widehat{P},\widehat{Q}) = \sup_{h_1,h_2\in\mathcal{H}} \Big| & \widehat{\mathbb{E}}_{x\sim\widehat{P}}\big[\mathbf{1}\{h_1(x)\neq h_2(x)\}\big] \\ & - \widehat{\mathbb{E}}_{x\sim\widehat{Q}}\big[\mathbf{1}\{h_1(x)\neq h_2(x)\}\big]\Big|.\end{aligned} \tag{E.1}$$

where $\widehat{\mathbb{E}}_{x\sim\widehat{P}}$ denotes the empirical expectation under sample $\widehat{P}$. This measures how well two empirical distributions can be separated by pairs of hypotheses in $\mathcal{H}$.

Definition 7 (Empirical Rademacher Complexity)**.** *Let $\widehat{S} = \{z_1,\ldots,z_n\}$ be an i.i.d. sample from some distribution over $\mathcal{X}$. Let $\mathcal{H}$ be a class of real-valued functions $h : \mathcal{X}\to\mathbb{R}$. The* empirical Rademacher complexity *of $\mathcal{H}$ on $\widehat{S}$ is*

$$\widehat{\mathfrak{R}}_{\widehat{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(z_i)\right], \tag{E.2}$$

where each σ_i is an independent uniform random variable taking values in $\{\pm1\}$. This complexity quantifies how richly $\mathcal{H}$ can fit random noise in the sample $\widehat{S}$.

Definition 8 (Alignment Error λ)**.** *Given two distributions P_r (retained) and P_f (forgotten), and a hypothesis class $\mathcal{H}$, define*

$$\lambda = \inf_{h^*\in\mathcal{H}}\Big[\epsilon_{P_r}(h^*) + \epsilon_{P_f}(h^*)\Big], \tag{E.3}$$

where $\epsilon_{P_r}(h^*)$ is the (true) expected classification error of h^* under P_r, and likewise $\epsilon_{P_f}(h^*)$ is the expected error under P_f. A smaller λ indicates that a single hypothesis can simultaneously fit both P_r and P_f well.

Appendix E.1. Formal Statement and Proof of Proposition 2 (Directional Role of σ)

Proposition 2 (Directional role of OOD intensity σ)**.** *Let P_f^0 denote the pre-unlearning representation distribution of the forget set and P_f^σ the post-unlearning distribution. Define the residual matching error $\eta_\sigma := \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^\sigma, Q_\sigma)$. Suppose the OOD family $\{Q_\sigma\}$ satisfies the monotonicity condition: for all $\sigma_2 > \sigma_1$,*

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_{\sigma_2}) \;\geq\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_{\sigma_1}).$$

This condition holds by construction for Gaussian noise families. Then the triangle inequality for $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ gives:

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^\sigma, P_f^0) \;\geq\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_\sigma) - \eta_\sigma.$$

Hence, under the monotonicity condition, increasing σ enlarges a lower bound on the displacement of forgotten representations (strengthening forgetting). At the same time, a larger $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q_\sigma)$ entering the retained-data risk bound of Proposition 1 can reduce retained-data utility, yielding the expected forgetting–utility tension.

The triangle inequality for $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ states that for any three distributions A, B, C:

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(A, C) \;\leq\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(A, B) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(B, C).$$

Applying this with $A = P_f^0$, $B = Q_\sigma$, $C = P_f^\sigma$:

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, P_f^\sigma) \;\leq\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_\sigma) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(Q_\sigma, P_f^\sigma).$$

Rearranging (and noting $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ is symmetric):

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^\sigma, P_f^0) \;\geq\; \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_\sigma) - \eta_\sigma.$$

Table C.1: Performance comparison of different machine unlearning methods across various datasets with single sample forgetting. Best results are highlighted in **bold**.

Dataset	Metric	AdaptForget	AFS	Amnesiac	Blindspot	FisherForgetting	INS	SSD-Tuning
	Acc	**0.7940**	0.7292	0.7435	0.7755	0.7375	0.3789	0.6116
	F1	**0.9920**	0.9370	0.7490	0.7750	0.7370	0.2800	0.5960
PathMNIST	IVD-E	**45.9820**	26.9050	14.5120	20.3750	21.5410	24.6890	19.8350
	IVD-M	**1049.3140**	634.1700	274.1710	430.2190	444.3910	516.2960	409.4190
	IVD-S	**547.6480**	330.5375	144.3415	225.2970	232.9660	270.4925	214.6270
	Acc	**0.7541**	0.7509	0.7272	0.7034	0.6483	0.5092	0.7379
	F1	**0.9091**	0.8864	0.6943	0.6782	0.6018	0.4462	0.7017
OCTMNIST	IVD-E	**24.7910**	23.9040	8.7780	12.7660	13.5820	13.5080	10.4020
	IVD-M	**532.0170**	525.0380	142.3720	249.4130	271.0390	240.0410	183.2170
	IVD-S	**278.4040**	274.4710	75.5750	131.0895	142.3105	126.7745	96.8095
	Acc	**0.5840**	0.5549	0.4479	0.4583	0.3535	0.4106	0.4704
	F1	**0.5016**	0.2018	0.4382	0.4443	0.3056	0.2836	0.3990
RetinaMNIST	IVD-E	**45.8870**	39.2630	34.0690	33.0690	37.6050	38.4180	18.6040
	IVD-M	**1022.0390**	906.6650	702.3710	708.8850	828.3660	820.9020	379.9700
	IVD-S	**533.9630**	472.9640	368.2200	370.9770	432.9855	429.6600	199.2870
	Acc	**0.7250**	0.7246	0.7097	0.7140	0.5558	0.3792	0.7214
	F1	**0.8252**	0.6026	0.7140	0.7185	0.5463	0.2197	0.6588
DermaMNIST	IVD-E	**49.1640**	37.1130	20.5000	25.2100	25.8520	34.3520	18.4670
	IVD-M	**1025.6750**	847.5190	451.4390	543.5100	558.0850	704.0240	401.6730
	IVD-S	**537.4195**	442.3160	235.9695	284.3600	291.9685	369.1880	210.0700
	Acc	**0.9536**	0.9367	0.8509	0.9453	0.9046	—	0.9529
	F1	**0.9672**	0.9554	0.8399	0.9452	0.8970	—	0.9522
ASD	IVD-E	**8.9410**	7.6520	3.0670	3.2590	3.4250	—	2.1680
	IVD-M	**53.2460**	51.5350	17.1160	18.5860	19.6960	—	11.8040
	IVD-S	**31.0935**	29.5935	10.0915	10.9225	11.5605	—	6.9860
	Acc	**0.7299**	0.7289	0.7254	0.7255	0.6307	—	0.7238
	F1	**0.7283**	0.7278	0.7248	0.7252	0.5874	—	0.7232
Diabetes	IVD-E	**2.3130**	1.0790	1.0860	0.9230	0.9400	—	0.5470
	IVD-M	**7.7490**	5.5550	4.1010	4.1640	4.3420	—	2.3850
	IVD-S	**5.0310**	3.3170	2.5935	2.5435	2.6410	—	1.4660
	Acc	**0.7315**	0.7123	0.6960	0.6813	0.6066	—	0.6941
	F1	**0.7352**	0.7209	0.6940	0.6763	0.5508	—	0.6919
MDD	IVD-E	**1.9880**	1.1550	1.0370	1.5560	1.6220	—	0.7030
	IVD-M	**9.0970**	6.1250	4.2420	7.8930	8.3700	—	3.4320
	IVD-S	**5.5425**	3.6400	2.6395	4.7245	4.9960	—	2.0675

Under the monotonicity condition, $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(P_f^0, Q_\sigma)$ is non-decreasing in σ, so the lower bound on the displacement grows. The retained-data utility claim follows directly from Proposition 1: since Q_σ enters the discrepancy term $|\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_r, Q_\sigma) - \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_f, Q_\sigma)|$ in the risk bound, a larger σ that simultaneously moves Q_σ away from P_r can loosen the bound and degrade retained-data utility. □

Note: This result provides a conditional, qualitative explanation of the directional role of σ; it is not a certified privacy guarantee and does not substitute for the empirical ablation in Section 4.3.

Appendix F. Proof of Proposition 1

Step 1: Decomposing the retained-domain error $\epsilon_{P_r}(h)$

We begin by defining the optimal hypothesis h^* that minimizes the classification error across both retained and forgotten data distributions:

$$\lambda = \inf_{h^* \in \mathcal{H}} \Big[\epsilon_{P_r}(h^*) + \epsilon_{P_f}(h^*)\Big]. \tag{F.1}$$

By adding and subtracting $\epsilon_{P_r}(h^*)$ and $\epsilon_{P_f}(h^*)$, we rewrite the classification error on P_r for any hypothesis $h \in \mathcal{H}$ as:

$$\begin{aligned}\epsilon_{P_r}(h) &= \epsilon_{P_r}(h^*) + (\epsilon_{P_r}(h) - \epsilon_{P_r}(h^*)) \\ &= \lambda + \Big[(\epsilon_{P_r}(h) - \epsilon_{P_r}(h^*)) + (\epsilon_{P_f}(h) - \epsilon_{P_f}(h^*))\Big].\end{aligned} \tag{F.2}$$

The alignment error λ represents the best achievable error trade-off between P_r and P_f. Our goal is now to bound the additional term involving $\epsilon_{P_r}(h)$ and $\epsilon_{P_f}(h)$.

Table C.2: Summary of Dataset Splits and Concepts

Abbreviation	Full Name	Description
QO	Query dataset overlapped with the training dataset	The original dataset used for querying the model. It includes data points that are part of the training set.
QNO	Query dataset disjoint with the training dataset	A dataset consisting of data points not seen by the model during training, used to query the model and assess its ability to forget.
QF	Query dataset	A subset of the original training dataset that has been explicitly targeted for forgetting by the unlearning algorithm.
OOD	Out-of-Distribution	A dataset that contains data points from a distribution different from the training data. It is used to improve the feature-level unlearning to forget.
Train set	Train Dataset	A training dataset is utilized to instruct a model on how to comprehend and predict input data. It comprises a variety of examples that enable the model to learn by adjusting its parameters, distinct from the test dataset which is solely used for evaluating the model's final performance.
Calibration	Calibration Dataset	A dataset that is disjoint from the training set but similar to the test set. It is used exclusively to calculate the p-value and evaluate the forgetting process.
Test Set	Test Dataset	A dataset used to evaluate the performance of the model after training. It is disjoint from the training data and used to assess metrics such as accuracy and F1-score.
k	Proportion k	The proportion of the training data used to train the student model in some unlearning methods, often represented as a fraction of the total training data.

Step 2: Relating P_r and P_f to the OOD domain Q

We assume that the model has no predictive capability on unseen OOD data, leading to the assumption:

$$\epsilon_Q(h) = \epsilon_Q(h^*). \tag{F.3}$$

Using domain adaptation bounds (cf. (Ben-David, 2010)), we express the classification error differences in terms of the $\mathcal{H}$-divergence between distributions. Specifically, for any pair of hypotheses (h_1, h_2), the following inequality holds:

$$\left|\epsilon_{P_f}(h_1, h_2) - \epsilon_Q(h_1, h_2)\right| \leq d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q), \tag{F.4}$$

and similarly for P_r. This allows us to rewrite $\epsilon_{P_f}(h)$ and $\epsilon_{P_r}(h)$ as:

$$\begin{aligned} \epsilon_{P_f}(h) &\leq \epsilon_Q(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q), \\ \epsilon_{P_r}(h) &\leq \epsilon_Q(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q). \end{aligned} \tag{F.5}$$

Substituting these bounds into the earlier decomposition of $\epsilon_{P_r}(h)$, we obtain:

$$\begin{aligned} \epsilon_{P_r}(h) \leq \lambda + \big[&(\epsilon_Q(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q) - \epsilon_Q(h)) \\ &+ (\epsilon_Q(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q) - \epsilon_Q(h))\big]. \end{aligned} \tag{F.6}$$

By assuming that the model exhibits no predictive advantage on the OOD data, we set $\epsilon_Q(h) = \epsilon_Q(h^*) = c_Q$, where c_Q is a constant independent of the hypothesis. This assumption implies that $\epsilon_Q(h) - \epsilon_Q(h^*) = 0$, leading to:

$$\epsilon_{P_r}(h) \leq \lambda + \big[d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q) - d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q)\big]. \tag{F.7}$$

Thus, the term $\big[d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q) - d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q)\big]$ naturally appears: a large $d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q)$ increases the error, while a small $d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q)$ helps remove P_f.

Step 3: Bounding the empirical risk using Rademacher complexity

We now switch to empirical estimates. The empirical risk $\hat{\epsilon}_S(h)$ provides an approximation to $\epsilon_{P_r}(h)$, which we bound using Rademacher complexity. A standard generalization bound for $\mathcal{H}$-bounded classification (cf. (Bartlett, 2002)) states that with probability at least $1 - \delta$:

$$\epsilon_{P_r}(h) \leq \hat{\epsilon}_S(h) + 2\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{F.8}$$

This follows from McDiarmid's inequality and uniform convergence bounds.

Similarly, the $\mathcal{H}\Delta\mathcal{H}$-divergence between two distributions is estimated from a finite sample. Instead of assuming direct approximations, we introduce a bound based on sample complexity (cf. (Mohri, 2018)):

$$\begin{aligned} \left|d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q) - \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_r, \hat{Q})\right| &\leq \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2m}}, \\ \left|d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q) - \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_f, \hat{Q})\right| &\leq \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2m}}. \end{aligned} \tag{F.9}$$

where m is the number of unlabeled samples from P_r and P_f used to estimate the discrepancy.

Step 4: Ensuring non-negativity and final bound derivation

Since $d_{\mathcal{H}\Delta\mathcal{H}}(P_r, Q) - d_{\mathcal{H}\Delta\mathcal{H}}(P_f, Q)$ may be negative in some cases, we take its absolute value to maintain a valid bound:

$$\left|\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_r, \hat{Q}) - \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_f, \hat{Q})\right|. \tag{F.10}$$

Using the above discrepancy bounds, we obtain the final generalization bound:

$$\epsilon_{P_r}(h) \le \hat{\epsilon}_S(h) + \left|\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_r, \hat{Q}) - \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}_f, \hat{Q})\right| + \lambda + 2\,\mathfrak{R}_n(\mathcal{H}) + 2\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{F.11}$$

This completes the proof. □

Appendix G. Convergence Analysis (Theorem 2)

Proof. Our goal is to show that minimizing the OOD-FOD loss ensures that the feature representations of the forgotten data P_f become indistinguishable from those of the OOD data Q in both feature and output spaces as training proceeds.

Step 1: Domain Discriminator and Optimal Solution

The OOD-FOD loss consists of two main components: (1) a domain adversarial term that enforces feature-level alignment and (2) a KL divergence term that enforces classifier output alignment. We first analyze the domain adversarial component.

A domain discriminator $D_\psi(z)$ is trained to distinguish between the feature representations of the forgotten data $P_f(z)$ and the OOD data $Q(z)$, where $z = F_\theta(x)$ is the extracted feature representation. The corresponding binary cross-entropy loss for the domain discriminator is:

$$\mathcal{L}_{\text{DA}} = -\mathbb{E}_{z\sim P_f}\big[\log D_\psi(z)\big] - \mathbb{E}_{z\sim Q}\big[\log(1 - D_\psi(z))\big]. \tag{G.1}$$

Following adversarial domain adaptation theory (Ganin, 2016), the optimal discriminator that minimizes this loss is given by:

$$D^*_\psi(z) = \frac{P_f(z)}{P_f(z) + Q(z)}. \tag{G.2}$$

Step 2: Substituting the Optimal Discriminator

Substituting $D^*_\psi(z)$ into the adversarial loss, we obtain:

$$\mathcal{L}^*_{\text{DA}} = -\mathbb{E}_{z\sim P_f}\left[\log \frac{P_f(z)}{P_f(z) + Q(z)}\right] - \mathbb{E}_{z\sim Q}\left[\log \frac{Q(z)}{P_f(z) + Q(z)}\right]. \tag{G.3}$$

Rewriting the expectations explicitly, we recognize that this corresponds to the Jensen-Shannon divergence (JSD) between $P_f(z)$ and $Q(z)$:

$$\mathcal{L}^*_{\text{DA}} = -\log 4 + 2\text{JSD}(P_f(z) \parallel Q(z)). \tag{G.4}$$

Thus, minimizing $\mathcal{L}_{\text{DA}}$ forces the feature distributions of forgotten data and OOD data to become similar, i.e.,

$$D_{\mathcal{H}\Delta\mathcal{H}}(P_f(F_\theta), Q(F_\theta)) \to 0, \quad \text{as} \quad t \to \infty. \tag{G.5}$$

This ensures that the forgotten data and OOD data are indistinguishable in feature space.

Step 3: Classifier Output Alignment and KL Divergence

The second term in OOD-FOD loss enforces classifier output alignment through the KL divergence:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{z_f\sim P_f, z_q\sim Q}\big[\text{KL}\big(C_\phi(z_f) \parallel C_\phi(z_q)\big)\big]. \tag{G.6}$$

Using Pinsker's inequality (Cover, 1999), we obtain:

$$d_{\text{TV}}(C_\phi(z_f), C_\phi(z_q)) \le \sqrt{\frac{1}{2}\text{KL}(C_\phi(z_f) \parallel C_\phi(z_q))}. \tag{G.7}$$

Since minimizing $\mathcal{L}_{\text{KL}}$ forces the KL divergence to approach zero, we obtain:

$$d_{\text{TV}}(P(C_\phi(F_\theta(x_f))), P(C_\phi(F_\theta(x_q)))) \le \epsilon(t), \quad \forall x_f \sim P_f, \quad x_q \sim Q. \tag{G.8}$$

This ensures that the classifier outputs for the forgotten data align with those of the OOD data.

By combining the results from Step 2 and Step 3, we conclude:

$$D_{\mathcal{H}\Delta\mathcal{H}}(P_f(F_\theta), Q(F_\theta)) \to 0, \quad d_{\text{TV}}(P(C_\phi(F_\theta(x_f))), P(C_\phi(F_\theta(x_q)))) \le \epsilon(t). \tag{G.9}$$

Thus, as $t \to \infty$, minimizing $\mathcal{L}_{\text{OOD-FOD}}$ guarantees that the feature representations and classifier predictions of the forgotten data P_f align with those of the OOD data Q, achieving structured forgetting.

Appendix H. Dataset preparation

Medical imaging datasets

In our benchmark experiments, we applied AdaptForget to four image datasets from MedMNIST (Yang, 2023): PathMNIST, RetinaMNIST, DermaMNIST, and OCTMNIST.

PathMNIST: This dataset consists of 100,000 non-overlapping image patches from hematoxylin and eosin-stained histological images, and 7,180 image patches from various clinical centers. It includes nine types of tissues: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and COAD epithelium. All images in PathMNIST are 3×28×28. To ensure a fair comparison and consistency with the experimental setup of AFS (Zhou, 2023) (applicable to other datasets as well), we randomly sampled 10,000 images for training, 5,000 images for testing, and 1,000 images, disjoint from the training dataset, for calibration.

RetinaMNIST: Derived from the DeepDRiD challenge (Liu, 2022), this dataset contains 1,600 retinal fundus images from 400 patients. These images are used for ordinal regression to grade the severity of diabetic retinopathy on a five-level scale. The original source images were 3×1,736×1, 824 pixels, which were center-cropped to the length of the shorter edge and resized to 3×28×28 pixels for standardization. We randomly sample 1,080 images for training, 120 images for testing, 400 images for validation, and 200 images disjoint from the training dataset for calibration.

DermaMNIST: Originating from HAM10000, this dataset is a comprehensive multi-source collection of dermatoscopic images representing common pigmented skin lesions. It includes 10,015 images categorized into seven distinct disease classes: actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc). For standardized analysis, images were resized from 3×600×450 pixels to 3×28×28 pixels. We randomly sample 7,007 images for training, 1,003 images for testing, 2,005 images for validation, and 1,000 images disjoint from the validation dataset for calibration.

OCTMNIST: This dataset, derived from an earlier dataset, contains 109,309 valid optical coherence tomography (OCT) images for diagnosing retinal diseases. The images are divided into four diagnostic categories: choroidal neovascularization, diabetic macular edema, drusen, and normal. The source images were in grayscale with sizes ranging from 384×277 to 1,536×512 pixels, which were center-cropped and resized to 1×28×28 pixels for uniformity. We randomly sample 90,000 images for training, 1,000 images for testing, 10,000 images for validation, and 1,000 images disjoint from the training dataset for calibration.

Clinical Health Record Datasets

For the application of AdaptForget in clinical tabular data, we used three datasets of clinical health records: Autism Spectrum Disorder (ASD), Mart for Intensive Care III (MIMIC) Death Dataset (MDD), and the Diabetes dataset.

ASD: This dataset for toddlers (Thabtah, 2017) contains 20 features from 1,054 records used to identify influential autistic traits and enhance the classification of ASD. We randomly sample 500 data points for training, 100 points for testing, and 100 points disjoint from the training dataset for calibration.

MDD: Sourced from the MIMIC dataset (Johnson, 2016), it includes 4,309 records with 57 features used to predict patient mortality, including information on rare diseases. We randomly sample 3,000 data points for training, 500 points for testing, 800 points for validation, and 400 points disjoint from the training dataset for calibration.

Diabetes: It contains responses from 441,455 individuals with 330 features, and was processed to comprise 49,485 records with 21 features for predicting diabetes incidence (Smith, 1988). We randomly sample 20,000 points for training, 5,000 points for testing, 2,000 points for validation, and 1,000 points disjoint from the validation dataset for calibration.

For the seven datasets, we randomly sampled partial data from the training dataset with percentages k from $\{0.25, 0.5, 0.75\}$ for AdaptForget. We prepare query datasets with different sizes N from $\{1, 10, 100, 500, 1000, 2000\}$. A query dataset that completely overlaps with the training dataset is labeled as QO, while a query dataset that is completely disjoint with the training dataset is labeled QNO. In particular, we emphasize the single-entry data forgetting setting for the query dataset designed to be forgotten by focusing on QF where $N = 1$. This setting is critical to demonstrate the effectiveness of AdaptForget in achieving reliable forgetting of single data points and complying with stringent data privacy regulations. See Table C.2 for the summary of dataset splits and related concepts.

References

Voigt, P., and von dem Bussche, A., 2017. The EU General Data Protection Regulation (GDPR): A Practical Guide, 1st ed. Springer, Cham. https://doi.org/10.1007/978-3-319-57959-7.

ENISA, 2023. Health Threat Landscape. http://www.enisa.europa.eu/publications/health-threat-landscape (accessed 26 September 2023).

Greer, S.L., Rozenblum, S., Fahy, N., Brooks, E., Jarman, H., de Ruijter, A., Palm, W., Wismar, M., et al., 2022. Everything you always wanted to know about European Union health policies but were afraid to ask, third, revised ed. European Observatory on Health Systems and Policies, World Health Organization Regional Office for Europe.

Security Intelligence, 2023. Cost of a data breach 2023: Healthcare industry impacts. https://securityintelligence.com/articles/cost-of-a-data-breach-healthcare-industry-impacts/ (accessed 26 September 2023).

Muthuppalaniappan, M., and Stevenson, K., 2021. Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health. Int. J. Qual. Health Care, 33(1), mzaa117. https://doi.org/10.1093/intqhc/mzaa117.

HIPAA Journal, 2023. Healthcare data breach statistics. https://www.hipaajournal.com/healthcare-data-breach-statistics/ (accessed 26 September 2023).

Lampropoulos, K., Zarras, A., Lakka, E., Barmpaki, P., Drakonakis, K., Athanatos, M., Debar, H., Alexopoulos, A., Sotiropoulos, A., Tsakirakis, G., et al., 2023. White paper on cybersecurity in the healthcare sector. The HEIR solution. arXiv preprint arXiv:2310.10139. https://doi.org/10.48550/arXiv.2310.10139

U.S. Congress, 1996. Health Insurance Portability and Accountability Act of 1996. Public Law 104–191, 110 Stat. 1936 (Aug. 21, 1996).

Fredrikson, M., Jha, S., and Ristenpart, T., 2015. Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), 1322–1333. https://doi.org/10.1145/2810103.2813677.

Ganju, K., Wang, Q., Yang, W., Gunter, C.A., and Borisov, N., 2018. Property inference attacks on fully connected neural networks using permutation invariant representations. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS), 619–633. https://doi.org/10.1145/3243734.3243834.

Nguyen, T.T., Huynh, T.T., Nguyen, P.L., Liew, A.W.-C., Yin, H., and Nguyen, Q.V.H., 2022. A survey of machine unlearning. arXiv preprint arXiv:2209.02299. https://doi.org/10.48550/arXiv.2209.02299.

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., and Johannes, R.S., 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care (SCAMC), 261.

Zhou, J., Li, H., Liao, X., Zhang, B., He, W., Li, Z., Zhou, L., and Gao, X., 2023. A unified method to revoke the private data of patients in intelligent healthcare with audit to forget. Nature Communications, 14, 6255. https://doi.org/10.1038/s41467-023-41703-x.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J.W., 2010. A theory of learning from different domains. Machine Learning, 79, 151–175. https://doi.org/10.1007/s10994-009-5152-4.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V., 2016. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N., 2021. Machine unlearning. 2021 IEEE Symposium on Security and Privacy (SP), 141–159. https://doi.org/10.1109/SP40001.2021.00019.

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., et al., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education. 23(1), 689. https://doi.org/10.1186/s12909-023-04698-z.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216.

Zhu, L., Liu, Z., and Han, S., 2019. Deep leakage from gradients. Advances in Neural Information Processing Systems. 32.

Dosovitskiy, A., and Brox, T., 2016. Inverting visual representations with convolutional networks. CVPR. 4829–4837. https://doi.org/10.1109/CVPR.2016.522.

Mahendran, A., and Vedaldi, A., 2015. Understanding deep image representations by inverting them. CVPR. 5188–5196. https://doi.org/10.1109/CVPR.2015.7299155.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M., 2020. Inverting gradients—how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems. 33, 16937–16947.

He, Z., Zhang, T., and Lee, R.B., 2019. Model inversion attacks against collaborative inference. ACSAC. 148–162. https://doi.org/10.1145/3359789.3359820.

Qu, Y., Yuan, X., Ding, M., Ni, W., Rakotoarivelo, T., and Smith, D., 2023. Learn to unlearn: A survey on machine unlearning. arXiv preprint arXiv:2305.07512. https://doi.org/10.48550/arXiv.2305.07512.

Nguyen, L., Rai, P., and Hoi, S.C.H., 2023. A survey of machine unlearning. arXiv preprint arXiv:2301.04220. https://doi.org/10.48550/arXiv.2209.02299.

Johnson, A.E.W., Pollard, T.J., Shen, L., et al., 2016. MIMIC-III, a freely accessible critical care database. Scientific Data. 3, 160035. https://doi.org/10.1038/sdata.2016.35.

Zhao, Z., Cao, L., and Lin, K.-Y., 2023. Supervision adaptation balancing in-distribution generalization and out-of-distribution detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 45(12), 15743–15758. https://doi.org/10.1109/TPAMI.2023.3286426.

Lu, W., Wang, J., Sun, X., Chen, Y., Ji, X., Yang, Q., and Xie, X., 2024. Diversify: A general framework for time series out-of-distribution detection and generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence. 46(6), 4534–4550. https://doi.org/10.1109/TPAMI.2023.3330037.

Yang, J., Zhou, K., Li, Y., and Liu, Z., 2024. Generalized out-of-distribution detection: A survey. International Journal of Computer Vision. 132(12), 5635–5662. https://doi.org/10.1007/s11263-024-02026-4.

He, R., Han, Z., Nie, X., Yin, Y., and Chang, X., 2024. Visual out-of-distribution detection in open-set noisy environments. International Journal of Computer Vision. 132(11), 5453–5470. https://doi.org/10.1007/s11263-024-02012-w.

Han, Z., Gui, X.-J., Sun, H., Yin, Y., and Li, S., 2022. Towards accurate and robust domain adaptation under multiple noisy environments. IEEE Transactions on Pattern Analysis and Machine Intelligence. 45(5), 6460–6479. https://doi.org/10.1109/TPAMI.2022.3182269.

Li, H., Wang, Y., Wan, R., Wang, S., Li, T.-Q., and Kot, A., 2020. Domain generalization for medical imaging classification with linear-dependency regularization. Advances in Neural Information Processing Systems. 33, 3118–3129.

Thota, M., and Leontidis, G., 2021. Contrastive domain adaptation. CVPR. 2209–2218. https://doi.org/10.1109/CVPR46437.2021.00225.

Liu, H., Wang, J., and Long, M., 2021. Cycle self-training for domain adaptation. Advances in Neural Information Processing Systems. 34, 22968–22981.

Guan, H., and Liu, M., 2021. Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering. 69(3), 1173–1185. https://doi.org/10.1109/TBME.2021.3092864.

Ganin, Y., and Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. ICML. 1180–1189.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T., 2018. Learning to generalize: meta-learning for domain generalization. AAAI. 32(1).

Li, H., Pan, S.J., Wang, S., and Kot, A.C., 2018. Domain generalization with adversarial feature learning. CVPR. 5400–5409. https://doi.org/10.1109/CVPR.2018.00565.

Han, Z., Zhang, Z., Wang, F., He, R., Su, W., Xi, X., and Yin, Y., 2023. Discriminability and transferability estimation: a Bayesian source importance estimation approach for multi-source-free domain adaptation. AAAI. 37(6), 7811–7820.

Han, Z., Sun, H., and Yin, Y., 2022. Learning transferable parameters for unsupervised domain adaptation. IEEE Transactions on Image Processing. 31, 6424–6439. https://doi.org/10.1109/TIP.2022.3197373.

Robey, A., Pappas, G.J., and Hassani, H., 2021. Model-based domain generalization. Advances in Neural Information Processing Systems. 34, 20210–20229.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C.C., 2022. Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 45(4), 4396–4415. https://doi.org/10.1109/TPAMI.2022.3159369.

Han, Z., Luo, G., Sun, H., Li, Y., Han, B., Gong, M., Zhang, K., and Liu, T., 2025. AlignCLIP: Navigating the misalignments for robust vision–language generalization. Machine Learning. 114(3), 1–19. https://doi.org/10.1007/s10994-025-06528-5.

Cover, T.M., and Thomas, J.A., 1999. Elements of Information Theory, 2nd ed. Wiley, New York.

Bartlett, P.L., and Mendelson, S., 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research. 3(Nov), 463–482.

Mohri, M., Rostamizadeh, A., and Talwalkar, A., 2018. Foundations of Machine Learning, 2nd ed. MIT Press, Cambridge, MA.

Serra, J., Suris, D., Miron, M., Karatzoglou, A., 2018. Overcoming catastrophic forgetting with hard attention to the task. In: International Conference on Machine Learning (ICML), PMLR, 4548–4557.

Baumhauer, T., Schöttle, P., Zeppelzauer, M., 2022. Machine unlearning: Linear filtration for logit-based classifiers. Mach. Learn. 111(9), 3203–3226. https://doi.org/10.1007/s10994-022-06178-9.

Graves, L., Nagisetty, V., Ganesh, V., 2021. Amnesiac machine learning. Proc. AAAI Conf. Artif. Intell. 35(13), 11516–11524.

Leino, K., Fredrikson, M., 2020. Stolen Memories: Leveraging model memorization for calibrated white-box membership inference. In: 29th USENIX Security Symposium (USENIX Security 20), 1605–1622.

Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), 3–18. https://doi.org/10.1109/SP.2017.41.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S., 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). 268–282. https://doi.org/10.1109/CSF.2018.00027

Schelter, S., Grafberger, S., and Dunning, T., 2021. HedgeCut: Maintaining randomised trees for low-latency machine unlearning. In: Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21). 1545–1557. https://doi.org/10.1145/3448016.3457239

Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al., 2022. DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. Patterns. 3(6), 100512. https://doi.org/10.1016/j.patter.2022.100512

Tomašev, N., Glorot, X., Rae, J.W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., et al., 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 572(7767), 116–119. https://doi.org/10.1038/s41586-019-1390-1.

Tarun, A.K., Chundawat, V.S., Mandal, M., Kankanhalli, M.S., 2024. Fast Yet Effective Machine Unlearning. IEEE Trans. Neural Netw. Learn. Syst. 35(9), 13046–13055. https://doi.org/10.1109/TNNLS.2023.3266233.

Huang, Y., Li, X., Li, K., 2021. EMA: Auditing Data Removal from Trained Models. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021. Lecture Notes in Computer Science, vol. 12905. Springer, 793–803. https://doi.org/10.1007/978-3-030-87240-3_76.

Tarun, A.K., Chundawat, V.S., Mandal, M., Kankanhalli, M.S., 2023. Deep regression unlearning. In: International Conference on Machine Learning (ICML 2023). Proceedings of Machine Learning Research, 202, 33921–33939.

Cha, S., Cho, S., Hwang, D., Lee, H., Moon, T., Lee, M., 2024. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. Proceedings of the AAAI Conference on Artificial Intelligence. 38(10), 11186–11194. https://doi.org/10.1609/aaai.v38i10.28996.

Foster, J., Schoepf, S., and Brintrup, A., 2024. Fast Machine Unlearning without Retraining through Selective Synaptic Dampening. Proceedings of the AAAI Conference on Artificial Intelligence. 38(11), 12043–12051. https://doi.org/10.1609/aaai.v38i11.29092.

Golatkar, A., Achille, A., Soatto, S., 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9304–9312. https://doi.org/10.1109/CVPR42600.2020.00932.

Goel, S., Prabhu, A., Sanyal, A., Lim, S.-N., Torr, P.H.S., Kumaraguru, P., 2022. Towards adversarial evaluations for inexact machine unlearning. arXiv preprint arXiv:2201.06640.

He, R., Han, Z., Lu, X., and Yin, Y., 2022. Safe-Student for Safe Deep Semi-Supervised Learning With Unseen-Class Unlabeled Data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 14585–14594. https://doi.org/10.1109/CVPR52688.2022.01418.

Gomes, E.D.C., Alberge, F., Duhamel, P., and Piantanida, P., 2022. Igeood: An Information Geometry Approach to Out-of-Distribution Detection. In: International Conference on Learning Representations (ICLR).

Jiang, D., Sun, S., and Yu, Y., 2022. Revisiting flow generative models for Out-of-distribution detection. In: International Conference on Learning Representations (ICLR).

Zhang, L., Goldstein, M., and Ranganath, R., 2021. Understanding failures in out-of-distribution detection with deep generative models. In: International Conference on Machine Learning (ICML). 12427–12436.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B., 2023. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. 10(1), 41. https://doi.org/10.1038/s41597-022-01721-8.

Thabtah, F., 2017. Autism spectrum disorder screening: Machine learning adaptation and DSM-5 fulfillment. In: Proceedings of the 1st International Conference on Medical and Health Informatics (ICMHI 2017). 1–6. https://doi.org/10.1145/3107514.3107515.

O'Shea, K., and Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458. https://doi.org/10.48550/arXiv.1511.08458.

Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., Gao, X., 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nature Communications. 15(1), 5649. https://doi.org/10.1038/s41467-024-50043-3.

Van de Ven, G.M., Siegelmann, H.T., Tolias, A.S., 2020. Brain-inspired replay for continual learning with artificial neural networks. Nature Communications. 11(1), 4069. https://doi.org/10.1038/s41467-020-17866-2.

Huang, Y., Li, X., Li, K., 2021. EMA: Auditing Data Removal from Trained Models. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021. Lecture Notes in Computer Science, vol. 12905. Springer, 793–803. https://doi.org/10.1007/978-3-030-87240-3_76.

Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing. 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077.

Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., Müller, K.-R., 2024. From clustering to cluster explanations via neural networks. IEEE Transactions on Neural Networks and Learning Systems. 35(2), 1926–1940. https://doi.org/10.1109/TNNLS.2022.3185901.

Thabtah, F., 2017. Autism spectrum disorder screening: Machine learning adaptation and DSM-5 fulfillment. In: Proceedings of the 1st International Conference on Medical and Health Informatics (ICMHI 2017). 1–6. https://doi.org/10.1145/3107514.3107515.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 618–626.

- We propose AdaptForget, a domain-adaptive feature-level unlearning framework.
- Out-of-distribution guidance enables effective disentanglement of forgotten data.
- Isolation verification distance provides interpretable auditability in feature space.
- AdaptForget supports real-time single-entry forgetting for patient data privacy.
- Extensive results on medical imaging and health records show state-of-the-art gains.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: