MIST: MINIMAL INTERVENTION SELECTIVE TOKEN UNLEARNING FOR CO-OCCURRING DISEASE RECOGNITION IN MULTI-LABEL CHEST X-RAYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective machine unlearning removes the influence of a target concept from a trained model without retraining, and is increasingly needed in medical imaging for privacy requests, annotation errors and evolving clinical standards. In multi-label chest X-ray recognition, the task is unusually hard because diseases co-occur and share visual evidence. We observe that existing methods sit at the two ends of a forgetting–retention trade-off: gradient ascent forgets the target but damages retained diseases, whereas distillation-based unlearning preserves retained diseases but barely forgets. We trace this to three causes: (i) disease evidence is entangled in shared representations, (ii) current methods act on parameters or global features and cannot localise what to forget, and (iii) nothing in their objectives protects the evidence that co-occurring diseases share. To address the above causes, we propose MIST, a novel token-level framework for Vision Transformers that is composed of three key components: (1) Disease-specific Token Selection (DTS) uses gradient-based attribution to find tokens that are informative for the target disease and not for retained diseases; (2) Selective Unlearning (SU) suppresses the target only through these tokens; and (3) Co-occurrence-Aware Retention (CAR) combines teacher–student distillation with a shared-token preservation loss on images where target and retained diseases appear together. On NIH ChestX-ray14 with patient-level five-fold cross-validation, MIST lowers Effusion AUROC from 0.870 to 0.403 while retained-disease AUROC falls by only 0.007 and co-occurrence AUROC by 0.008, which is 65% and 69% less collateral degradation than SalUn at slightly stronger forgetting, and it modifies 37% of patch tokens. Our results demonstrate that MIST is the only method ranked in the top two on all six forgetting and retention measures.

1 INTRODUCTION

Models trained on medical images increasingly need a mechanism for removing what they have learned about specific data or concepts, because of privacy regulation (Mantelero, 2013), annotation errors, bias correction (Seyyed-Kalantari et al., 2021) and evolving clinical standards (Cao & Yang, 2015; Ginart et al., 2019). Machine unlearning aims to remove such influence without retraining from scratch, and recent methods show promising results on natural images (Golatkar et al., 2020; Chundawat et al., 2023; Kurmanji et al., 2023; Jia et al., 2023; Fan et al., 2024; Foster et al., 2024; Trippa et al., 2024); see Nguyen et al. (2022) and Xu et al. (2023) for surveys.

Despite this progress, selective unlearning in multi-label chest X-ray recognition remains difficult. Diseases frequently co-occur and share anatomical and visual evidence (Wang et al., 2017; Irvin et al., 2019; Johnson et al., 2019), and label co-occurrence is strong enough that classifiers are designed to exploit it (Chen et al., 2020; Guan & Huang, 2020): pleural effusion commonly appears with infiltration and cardiomegaly, so a Vision Transformer (ViT) (Dosovitskiy et al., 2021) learns representations in which these findings are entangled. Forgetting one disease can then degrade the recognition of the others, and in a clinical setting even small degradations of retained findings matter.

In our study, we find that existing methods sit at the two ends of a forgetting–retention trade-off (Figure 1b): gradient ascent (Trippa et al., 2024) forgets the target strongly but degrades retained

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

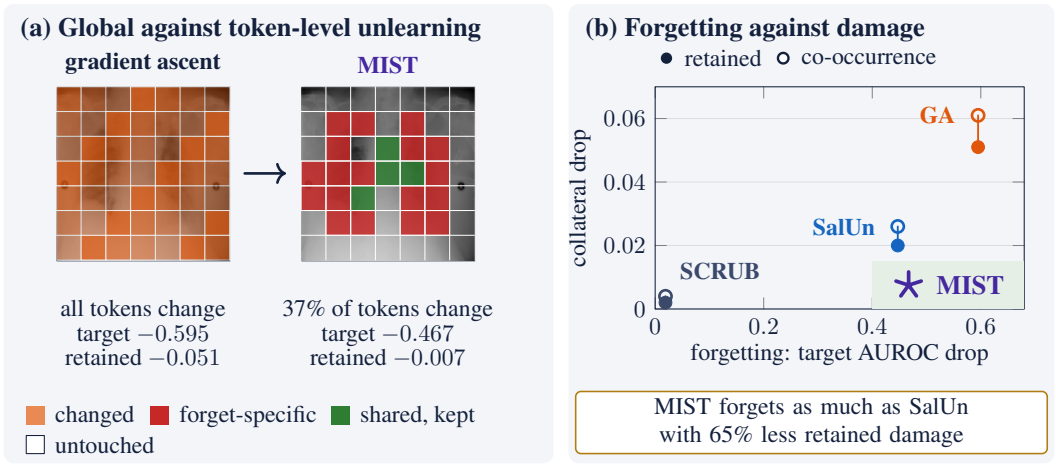


Figure 1: Motivation: (a) global unlearning against token-level unlearning; (b) forgetting–retention trade-off on NIH ChestX-ray14.

diseases, whereas SCRUB (Kurmanji et al., 2023) preserves retained diseases but barely forgets. We identify three causes. ① *Entangled evidence*: co-occurring diseases share image regions, so any change that removes the target also touches features the retained diseases rely on. ② *No localisation*: gradient-ascent, distillation-based and saliency-guided methods (Trippa et al., 2024; Kurmanji et al., 2023; Fan et al., 2024) act on parameters or global features, and cannot restrict forgetting to where the target-specific evidence lies. ③ *Unprotected shared evidence*: their objectives retain performance on average but place no constraint on the regions that target and retained diseases share, which is where interference arises.

To address these challenges, we propose MIST (Minimal Intervention Selective Token Unlearning), which operates on the patch tokens of a ViT and pairs each cause with one component. Disease-specific Token Selection (DTS) computes gradient-based attribution for the target and the retained diseases and selects tokens that are informative for the former and not for the latter. Selective Unlearning (SU) suppresses the target disease through these tokens only. Co-occurrence-Aware Retention (CAR) keeps the student close to a frozen teacher on retained data and, on images where target and retained diseases appear together, on the tokens they share.

Our contributions are summarised as follows:

- We characterise selective disease unlearning under label co-occurrence and show that parameter-level and feature-level methods occupy the extremes of the forgetting–retention trade-off on multi-label chest X-rays.
- We propose MIST, a token-level unlearning framework that pairs each identified cause with one mechanism: DTS for localisation, SU for target suppression, and CAR for the protection of shared evidence.
- On NIH ChestX-ray14 with patient-level five-fold cross-validation, MIST reduces Effusion AUROC by 0.467 with a retained-disease AUROC drop of 0.007 and a co-occurrence drop of 0.008, which is 65% and 69% lower than SalUn at comparable forgetting and 7× lower than gradient ascent, while modifying 37% of patch tokens; the behaviour holds across all five forget targets.

2 Related Work

2.1 Machine Unlearning

Machine unlearning was introduced to remove the influence of training data without full retraining (Cao & Yang, 2015), motivated by the right to be forgotten (Mantelero, 2013; Ginart et al., 2019). **Exact** methods such as SISA partition the data into shards and retrain only the affected sub-models (Bourtoule et al., 2021), and certified approaches bound the difference from a retrained model for convex or linearised settings (Guo et al., 2020; Sekhari et al., 2021; Izzo et al., 2021).

Approximate methods dominate for deep networks. Gradient-based approaches revise the model with updates computed on the forget set, by gradient ascent (Trippa et al., 2024), by approximately reversing the training updates (Thudi et al., 2022) or by undoing the batches that contained the forgotten data (Graves et al., 2021), and Descent-to-Delete studies gradient-based deletion with guarantees (Neel et al., 2021). Fisher-information methods scrub the weights that carry most information about the forget data (Golatkar et al., 2020; 2021), selective synaptic dampening attenuates forget-specific parameters without retraining (Foster et al., 2024), boundary unlearning shifts the decision boundary of the forget samples (Chen et al., 2023), and influence functions approximate the effect of removing individual samples (Koh & Liang, 2017; Warnecke et al., 2023). Distillation-based methods use a teacher–student objective (Hinton et al., 2015; Romero et al., 2015): a competent and an incompetent teacher in Chundawat et al. (2023), error-maximising noise with repair in Tarun et al. (2023), and a teacher that the student follows on the retain set and departs from on the forget set in SCRUB (Kurmanji et al., 2023). SalUn restricts updates to weights that are salient for the forget set (Fan et al., 2024), and Jia et al. (2023) show that sparsity narrows the gap to retraining. Unlearning can itself leak information through the difference between the two models (Chen et al., 2021; Shokri et al., 2017). All of these methods act on parameters or global features; none preserves the representations that correlated classes share, which is the difficulty in multi-label medical data. Retention under continued optimisation is also studied in continual learning (Kirkpatrick et al., 2017; Li & Hoiem, 2018), whose distillation-style constraints we adopt for retention.

2.2 Unlearning and Representation Removal in Medical Imaging

Unlearning in medical imaging is comparatively unexplored. Nasirigerdeh et al. (2024) provide one of the first systematic evaluations and show that existing methods can harm generalisation and retained clinical performance, particularly for larger forget sets. Forget-MI (Hardan et al., 2025) removes unimodal and joint patient representations in multimodal healthcare data with dedicated loss and perturbation strategies, Xue et al. (2024) remove undesirable data influence in MRI reconstruction while reducing hallucination artefacts, and Deng et al. (2024) combine feature-level forgetting with memory preservation for federated client unlearning. Falcao & Cordeiro (2026) show that unlearning may increase clinically harmful false negatives even when forgetting succeeds. Related work removes bias (Ashraf et al., 2018) and spurious correlations (Hagos et al., 2023) from chest X-ray models, which are known to rely on shortcuts, dataset-specific confounders and hidden stratification (Zech et al., 2018; Oakden-Rayner et al., 2020). These studies address sample-level removal, multimodal data, federated privacy or global feature modification. Selective forgetting of one disease among co-occurring diseases in multi-label chest X-ray recognition (Wang et al., 2017; Rajpurkar et al., 2017; Baltruschat et al., 2019) remains largely unexplored.

2.3 Token Representations and Attribution in Vision Transformers

Vision Transformers encode localised patch-level information (Dosovitskiy et al., 2021; Touvron et al., 2021), self-supervised variants produce tokens that align with semantic regions (Caron et al., 2021; Oquab et al., 2024), and such backbones are now standard in medical imaging (Hatamizadeh et al., 2022; Pérez-García et al., 2025). Token importance has been estimated with attention flow (Abnar & Zuidema, 2020), relevance propagation (Chefer et al., 2021) and gradient-weighted activations in the spirit of Grad-CAM and integrated gradients (Selvaraju et al., 2017; Sundararajan et al., 2017), and token-level selection is widely used to make ViTs efficient (Rao et al., 2021; Liang et al., 2022; Bolya et al., 2023). Tokens are therefore a natural unit for localised forgetting, but correlated diseases activate overlapping tokens, so removing disease-specific knowledge can suppress shared clinical semantics. Existing unlearning approaches do not localise forgetting within transformer token representations, which is the gap MIST addresses.

3 Methodology

Overview. As shown in Figure 2, MIST consists of three components. Disease-specific Token Selection (DTS, Section 3.2) uses a frozen teacher to locate the patch tokens that carry evidence for the target disease and not for the retained diseases. Selective Unlearning (SU, Section 3.3) suppresses the target disease in a trainable student through these tokens. Co-occurrence-Aware Retention (CAR, Section 3.4) keeps the student aligned with the teacher on retained data and on the tokens that the

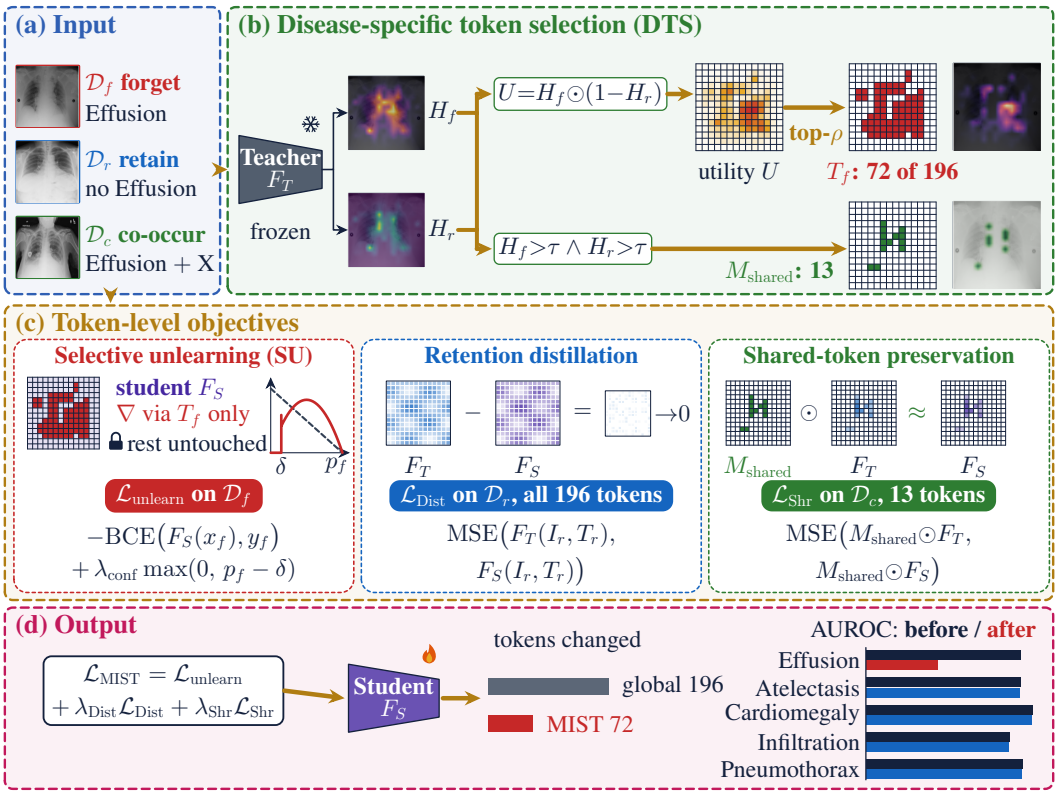


Figure 2: Overview of MIST: (a) input sets, (b) disease-specific token selection, (c) token-level objectives, (d) output.

target and retained diseases share. Unlike feature-level unlearning, which modifies parameters or representations globally, MIST confines the intervention to a sparse, disease-specific part of the token space and leaves the remainder of the representation untouched.

3.1 Preliminaries

Problem formulation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N_\mathcal{D}}$ be a multi-label chest X-ray dataset, where x_i is an image and $y_i \in \{0,1\}^C$ is the annotation vector over C diseases. Given a target forget disease d_f, the objective is to suppress the representations associated with d_f while maintaining performance on the retained diseases d_k, $k \neq f$. We partition the data into a forget set of samples associated with the target, a retain set of samples without it, and a co-occurrence set in which the target and a retained disease appear together:

$$\mathcal{D}_f = \{x_i \mid y_i^{(d_f)} = 1\}, \qquad \mathcal{D}_r = \{x_i \mid y_i^{(d_f)} = 0\}, \qquad \mathcal{D}_c = \{x_i \mid y_i^{(d_f)} = 1 \wedge y_i^{(d_k)} = 1\}. \quad (1)$$

Note that $\mathcal{D}_c \subseteq \mathcal{D}_f$: co-occurrence samples are exactly the forget samples on which a retained disease must still be recognised, and they are therefore where forgetting and retention conflict directly.

Teacher and student. A pretrained Rad-DINO Vision Transformer (Pérez-García et al., 2025), fine-tuned for multi-label classification, serves as the frozen teacher F_T. A trainable student F_S is initialised from the teacher weights and is the model that unlearns. For an input image, the encoder produces patch tokens $Z = \{z_1, \ldots, z_N\}$ with $z_i \in \mathbb{R}^{C_z}$ and $N = 196$ for 224×224 inputs, from which the classification head produces one logit s_d per disease. We write $F(I, T)$ for the token representations that a model F produces for images I on the token set T.

Why global unlearning interferes. Consider a generic unlearning loss $\mathcal{L}$ evaluated on a forget sample. By the chain rule, its effect on the encoder parameters θ passes through every token,

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial z_i} \frac{\partial z_i}{\partial \theta} = \underbrace{\sum_{i \in T_f} \frac{\partial \mathcal{L}}{\partial z_i} \frac{\partial z_i}{\partial \theta}}_{\text{target-specific}} + \underbrace{\sum_{i \in T_s} \frac{\partial \mathcal{L}}{\partial z_i} \frac{\partial z_i}{\partial \theta}}_{\text{shared}} + \underbrace{\sum_{i \notin T_f \cup T_s} \frac{\partial \mathcal{L}}{\partial z_i} \frac{\partial z_i}{\partial \theta}}_{\text{unrelated}}, \tag{2}$$

where T_f denotes tokens that carry target-specific evidence and T_s tokens whose evidence is shared with a retained disease. Parameter-level and feature-level methods optimise all three terms together. The second term is the source of collateral degradation on co-occurring diseases, and the third changes representations that have nothing to do with the target. MIST is designed to act through the first term, to constrain the second, and to leave the third untouched.

3.2 Disease-specific Token Selection

Observation 1. *Global unlearning trades forgetting against retention.* Under identical splits, initialisation and optimisation settings (Table 2), gradient ascent lowers the target AUROC by 0.595 but lowers retained-disease AUROC by 0.051 and co-occurrence AUROC by 0.061, while SCRUB keeps both within 0.004 but lowers the target AUROC by only 0.019. Neither method can separate target-specific evidence from shared evidence, because both act on the whole representation (causes ① and ②). The module therefore first localises the transformer representations associated with the target disease, so that the entire representation space need not be modified.

Disease-aware token attribution. Given the teacher F_T, disease-aware attribution maps are generated by gradient-weighted token-activation importance, in the spirit of gradient-based localisation (Selvaraju et al., 2017; Chefer et al., 2021). A first-order expansion of the disease logit around the token embeddings gives $s_d(Z) \approx s_d(\mathbf{0}) + \sum_{i=1}^{N} \sum_{c=1}^{C_z} \frac{\partial s_d}{\partial z_{i,c}} z_{i,c}$, so the inner sum is the first-order contribution of token i to the prediction of disease d. We take its magnitude and normalise over tokens:

$$H_d(i) = \frac{\left| \sum_{c=1}^{C_z} \frac{\partial s_d}{\partial z_{i,c}} z_{i,c} \right|}{\max_j \left| \sum_{c=1}^{C_z} \frac{\partial s_d}{\partial z_{j,c}} z_{j,c} \right| + \epsilon}, \tag{3}$$

where $\frac{\partial s_d}{\partial z_{i,c}}$ is the gradient of the disease logit with respect to feature dimension c of token i and ϵ is a small constant for numerical stability. The normalisation maps every attribution to $[0, 1]$ and makes maps of different diseases and images comparable, which the next step requires. The formulation highlights tokens that contribute strongly to a disease prediction and suppresses low-contribution regions. We compute H_f for the target forget disease and H_r for the retained diseases; both localise clinically informative regions (Figure 3).

Minimal-intervention mask. To suppress the target while preserving retained clinical semantics, MIST constructs a minimal-intervention utility map and selects a sparse token subset from it:

$$U = H_f \odot (1 - H_r), \qquad T_f = \mathrm{TopK}(U, \rho), \tag{4}$$

where $\odot$ denotes element-wise multiplication and ρ is the minimum information-coverage threshold for token selection. The utility of a token is high only if it is informative for the target *and* uninformative for the retained diseases, and it vanishes as $H_r(i) \to 1$ however informative the token is for the target. Table 1 lists the four regimes. Because both maps lie in $[0, 1]$, a token that the shared mask of Section 3.4 marks as shared ($H_r(i) > \tau$) has utility at most $(1 - \tau)H_f(i)$, so shared evidence is systematically ranked below target-specific evidence of the same strength. By selecting only a sparse subset of tokens, 72 of 196 (37%) for the full model, MIST forgets locally with minimal intervention on the original representation space.

3.3 Selective Unlearning

Motivation. The standard way to forget a class is to maximise its classification loss on the forget set. For a forget sample with $y^{(d_f)} = 1$, the term of $-\mathrm{BCE}$ that involves the target is $\log p_f$, where

Table 1: Token regimes induced by the two attribution maps.

$H_f(i)$	$H_r(i)$	Utility $U(i)$	Interpretation	Treatment in MIST
high	low	high	target-specific evidence	selected in T_f, suppressed (Section 3.3)
high	high	low	evidence shared with a retained disease	in M_{shared}, preserved (Section 3.4)
low	high	near zero	retained-disease evidence	untouched, distilled on $\mathcal{D}_r$
low	low	near zero	background	untouched

$p_f = \sigma(s_{d_f})$ is the predicted probability of the target. Its gradient with respect to the logit is

$$\frac{\partial \log p_f}{\partial s_{d_f}} = 1 - p_f, \tag{5}$$

which does not vanish as forgetting proceeds but tends to its maximum as $p_f \to 0$. An ascent objective therefore keeps pushing, with increasing force, after the target is already suppressed, which adds representation drift without adding forgetting. The objective should instead be bounded by an explicit suppression target.

Technique. The selective unlearning module suppresses target-disease representations using token-level forgetting objectives applied to the selected token subset T_f. Let (x_f, y_f) denote forget-set samples and labels. The forgetting objective maximises the prediction loss associated with the target disease,

$$\mathcal{L}_{\text{forget}} = -\text{BCE}\big(F_S(x_f), y_f\big), \tag{6}$$

and an additional confidence-suppression loss further suppresses residual target confidence,

$$\mathcal{L}_{\text{conf}} = \max(0,\, p_f - \delta), \qquad \mathcal{L}_{\text{unlearn}} = \mathcal{L}_{\text{forget}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}, \tag{7}$$

where δ is the target suppression threshold and λ_{conf} controls the contribution of confidence suppression. In contrast to Equation (5), the hinge is active only while $p_f > \delta$: its gradient with respect to the logit is $p_f(1 - p_f)$ above the threshold and exactly zero once $p_f \le \delta$, so it supplies a direct push on the target probability that switches itself off when the suppression target is met. The sensitivity analysis in Section 4.4 shows that forgetting grows steadily with λ_{conf}.

3.4 Co-occurrence-Aware Retention

Motivation. While the target is suppressed, the knowledge of retained and co-occurring diseases must be preserved. Average retention on $\mathcal{D}_r$ does not protect the images on which interference is most likely, namely those in $\mathcal{D}_c$, where the target and a retained disease share image regions (cause ③); these images belong to the forget set and are therefore pushed by $\mathcal{L}_{\text{unlearn}}$. The retention module reduces representation drift between the frozen teacher and the selectively unlearned student both on retained data in general and on the shared tokens of co-occurrence samples.

Co-occurrence-aware retention distillation. To preserve retained-disease semantics, MIST uses teacher–student representation distillation (Hinton et al., 2015; Romero et al., 2015) on retain-set samples. Let $F_T(I_r, T_r)$ and $F_S(I_r, T_r)$ denote the teacher and student retain representations. The retention distillation objective is

$$\mathcal{L}_{\text{Dist}} = \text{MSE}\big(F_T(I_r, T_r),\, F_S(I_r, T_r)\big), \tag{8}$$

which encourages the student to keep the retained-disease representations of the teacher during selective forgetting. Matching representations rather than predictions constrains the token space directly, which is where the intervention takes place.

Shared-token preservation. Co-occurring thoracic diseases frequently share clinically informative transformer regions. To preserve them, MIST constructs a shared-token mask from the overlap of the two attribution responses,

$$M_{\text{shared}} = \mathbf{1}(H_f > \tau) \odot \mathbf{1}(H_r > \tau), \tag{9}$$

where τ is the attribution threshold and $\mathbf{1}(\cdot)$ is the binary indicator function. A token enters the mask only if it is informative for *both* diseases, which is the second row of Table 1. Using the mask, the shared preservation loss on co-occurrence samples (I_c, T_c) is

$$\mathcal{L}_{\text{Shr}} = \text{MSE}\big(M_{\text{shared}} \odot F_T(I_c, T_c),\; M_{\text{shared}} \odot F_S(I_c, T_c)\big). \tag{10}$$

This objective constrains representation drift within the clinically shared regions and reduces interference with retained co-occurring diseases. It acts on precisely the second term of Equation (2), which no existing objective constrains.

3.5 Overall Objective

The overall objective combines selective forgetting, confidence suppression, retention distillation and shared-token preservation:

$$\mathcal{L}_{\text{MIST}} = \mathcal{L}_{\text{unlearn}} + \lambda_{\text{Dist}}\mathcal{L}_{\text{Dist}} + \lambda_{\text{Shr}}\mathcal{L}_{\text{Shr}}, \tag{11}$$

where λ_{Dist} and λ_{Shr} control the contributions of retention distillation and shared-token preservation. The three terms correspond to the three terms of Equation (2): $\mathcal{L}_{\text{unlearn}}$ acts through the target-specific tokens, $\mathcal{L}_{\text{Shr}}$ anchors the shared tokens, and $\mathcal{L}_{\text{Dist}}$ anchors the retained representation as a whole. The attribution maps are computed with the frozen teacher and require one backward pass per disease, and the student is updated for only five epochs (Section 4), so MIST avoids retraining. Algorithm 1 in Section A summarises the procedure.

4 Experiments

4.1 Settings

Data and protocol. We use NIH ChestX-ray14 (Wang et al., 2017) and, following prior multi-label chest X-ray studies, five diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration and Pneumothorax. To prevent patient-level leakage, splits are made by patient identifier with multi-label stratification (MultilabelStratifiedKFold (Sechidis et al., 2011)), so that no patient appears in more than one split and co-occurrence statistics are preserved. We use five-fold patient-level cross-validation; within each fold, 10% of training patients are held out for validation. Images without the selected labels or with missing files are excluded, and all images are resized to 224×224. Effusion is the primary forget target, and the other four diseases are retained; Section 4.3 rotates the target. These five diseases exhibit distinct co-occurrence patterns and varying degrees of anatomical overlap, which makes them suitable for evaluating selective unlearning. The co-occurrence subset $\mathcal{D}_c$ is used specifically to evaluate representation interference and the preservation of shared clinical semantics.

Baselines and metrics. We compare with gradient ascent (GA) (Trippa et al., 2024), SCRUB (Kurmanji et al., 2023) and SalUn (Fan et al., 2024), which represent global-optimisation, distillation-based and saliency-guided unlearning. All methods use the same splits, teacher initialisation, optimisation settings and evaluation protocol. Forgetting is measured by the drop in AUROC and average precision (AP) on the target disease (F-AUROC, F-AP; higher is stronger forgetting). Retention is measured by the mean drop over retained diseases (R-AUROC, R-AP) and by the drop on the co-occurrence subset $\mathcal{D}_c$ (C-AUROC, C-AP); lower is better.

Implementation. Experiments run in PyTorch (Paszke et al., 2019) on a single NVIDIA GeForce RTX 4090 (24 GB VRAM) with 64 GB of system RAM. The Rad-DINO backbone is fine-tuned for 50 epochs on the full training data and frozen as the teacher F_T for all unlearning experiments; the student F_S is initialised from the same weights. The framework operates on the 196 patch tokens produced by the transformer encoder, and the disease-aware attribution maps are computed from the teacher's predictions. The student is optimised with AdamW (Loshchilov & Hutter, 2019) (learning rate 1.5×10^{-4}, weight decay 0.05) for 5 epochs with batch size 32. Confidence suppression uses a target threshold of $\delta = 0.15$. The attribution threshold τ was set empirically to 0.55 to preserve shared tokens, and the information-coverage parameter ρ controls the sparsity of the selected forget tokens, giving about 37% token utilisation during selective forgetting.

4.2 Main Results

As shown in Table 2, the baselines occupy the ends of the trade-off. GA forgets most (F-AUROC 0.595) but lowers retained AUROC by 0.051 and co-occurrence AUROC by 0.061. SCRUB preserves best (0.002 and 0.004) but exhibits weak forgetting (0.019). MIST forgets more than SalUn (0.467 against 0.447) while its collateral degradation is 65% lower on retained AUROC (0.007 against

Table 2: Comparison of selective unlearning methods on NIH ChestX-ray14 (forget target: Effusion; mean ± standard deviation over five folds). F: drop on the forget target (higher is stronger forgetting); R and C: drop on retained diseases and on the co-occurrence subset (lower is better). MIST is the only method in the top two on all six measures: it forgets more than SalUn with a 65% lower retained AUROC drop.

Method	F-AUROC ↑	F-AP ↑	R-AUROC ↓	R-AP ↓	C-AUROC ↓	C-AP ↓
GA (Trippa et al., 2024)	0.595 ± .015	0.521 ± .014	0.051 ± .006	0.068 ± .007	0.061 ± .007	0.074 ± .008
SCRUB (Kurmanji et al., 2023)	0.019 ± .003	0.032 ± .004	0.002 ± .001	0.003 ± .001	0.004 ± .001	0.005 ± .001
SalUn (Fan et al., 2024)	0.447 ± .013	0.472 ± .014	0.020 ± .003	0.027 ± .004	0.026 ± .004	0.031 ± .004
MIST (ours)	0.467 ± .012	0.492 ± .013	0.007 ± .002	0.005 ± .001	0.008 ± .002	0.007 ± .002

Table 3: Per-disease AUROC and AP of the teacher and of the student before and after unlearning with MIST (mean ± standard deviation over five folds). Only the forget target changes materially: Effusion AUROC falls from 0.870 to 0.403, while each retained disease loses 0.007 AUROC.

	Teacher		Student before forget		Student after forget	
Disease	AUROC	AP	AUROC	AP	AUROC	AP
Atelectasis	0.871 ± .005	0.839 ± .004	0.871 ± .005	0.839 ± .004	0.864 ± .004	0.834 ± .005
Cardiomegaly	0.940 ± .004	0.919 ± .004	0.940 ± .004	0.919 ± .004	0.933 ± .005	0.914 ± .005
Effusion (forget)	0.870 ± .005	0.849 ± .005	0.870 ± .005	0.849 ± .005	0.403 ± .012	0.357 ± .011
Infiltration	0.811 ± .004	0.783 ± .005	0.811 ± .004	0.783 ± .005	0.804 ± .004	0.778 ± .004
Pneumothorax	0.882 ± .005	0.823 ± .004	0.882 ± .005	0.823 ± .004	0.875 ± .005	0.818 ± .004

0.020), 81% lower on retained AP, 69% lower on co-occurrence AUROC (0.008 against 0.026) and 77% lower on co-occurrence AP. Relative to GA, MIST gives up 0.128 of forgetting and reduces retained degradation sevenfold. MIST is the only method ranked in the top two on all six measures. Unlike GA and SCRUB, which update model representations globally, and SalUn, which performs parameter-level selective forgetting through saliency-guided weight masking, MIST performs localised token-level intervention and modifies only about 37% of the transformer patch tokens, which reduces unnecessary representation drift and improves the stability of retained diseases. Other medical imaging and optimisation-based unlearning frameworks target multimodal or data-removal settings rather than selective disease-level forgetting in multi-label classification, so the comparison uses the three methods that apply directly to this setting.

Table 3 reports per-disease performance before and after unlearning. Effusion AUROC falls from 0.870 to 0.403 and AP from 0.849 to 0.357, while Atelectasis, Cardiomegaly, Infiltration and Pneumothorax each lose 0.007 AUROC and at most 0.005 AP, including the diseases that co-occur most with Effusion.

4.3 Generalisation Across Forget Targets

Table 4 rotates the forget target over all five diseases. Forgetting remains strong (AUROC drop 0.398 to 0.467) and the mean retained drop stays between 0.006 and 0.009. Pneumothorax, whose evidence is spatially localised, is forgotten with the least interference (0.006). Cardiomegaly and Infiltration, whose evidence overlaps more with other thoracic findings, show slightly weaker forgetting and slightly larger retained drift, because more of their informative tokens are shared and are therefore protected. The strength of forgetting thus varies with the amount of shared evidence, as the design intends.

4.4 Ablation

Table 5 adds the components in turn. The forgetting loss alone forgets little (F-AUROC 0.178). Adding confidence suppression raises forgetting to 0.297 at $\lambda_{\text{conf}} = 0.05$ and 0.360 at $\lambda_{\text{conf}} = 0.10$. Adding retention distillation halves the retained drop (0.018 to 0.009) at some cost in forgetting (0.297 to 0.254). The full model, with shared-token preservation and $\lambda_{\text{conf}} = 0.15$, reaches the strongest forgetting (0.467) together with the lowest retained drop (0.007), and selects 72 tokens (37%) against 55 (28%) for the other configurations.

Table 4: MIST across forget targets (drop in AUROC and AP; mean ± std over five folds). Every target is forgotten by at least 0.398 AUROC while the mean retained drop stays at or below 0.009.

	Forget drop ↑		Mean retained drop ↓	
Forget disease	AUROC	AP	AUROC	AP
Effusion	0.467 ± .012	0.492 ± .013	0.007 ± .002	0.005 ± .001
Atelectasis	0.431 ± .011	0.455 ± .012	0.008 ± .002	0.007 ± .002
Cardiomegaly	0.412 ± .010	0.438 ± .011	0.009 ± .002	0.008 ± .002
Infiltration	0.398 ± .010	0.427 ± .011	0.009 ± .002	0.008 ± .002
Pneumothorax	0.438 ± .011	0.461 ± .012	0.006 ± .002	0.006 ± .002

Table 5: Component ablation and sensitivity to λ_{conf} (forget target: Effusion). Confidence suppression raises forgetting and distillation lowers the retained drop; only the full model attains both the strongest forgetting (0.467) and the lowest retained drop (0.007).

Configuration	λ_{conf}	Tokens	F-AUROC ↑	R-AUROC ↓	Token ratio
$\mathcal{L}_{\text{forget}}$ only	0.00	55	0.178	0.018	0.28
$\mathcal{L}_{\text{unlearn}}$	0.05	55	0.297	0.018	0.28
$\mathcal{L}_{\text{unlearn}}$	0.10	55	0.360	0.010	0.28
$\mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{Dist}}$	0.05	55	0.254	0.009	0.28
Full MIST ($\mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{Dist}} + \mathcal{L}_{\text{Shr}}$)	0.15	72	0.467	0.007	0.37

4.5 Qualitative Analysis

Figure 3 shows three Effusion co-occurrence cases. The attribution maps localise the evidence for Effusion and for the retained disease; the shared-token maps mark regions activated by both, which MIST preserves; and the suppressed-token maps show that the intervention is confined to Effusion-specific regions and does not modify the representation globally.

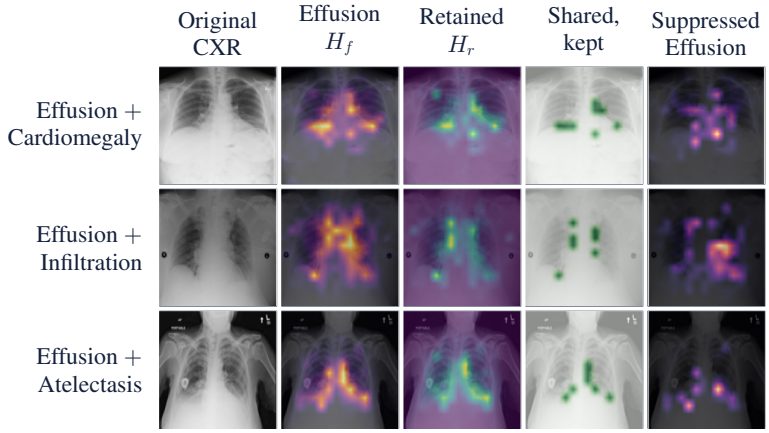


Figure 3: Token-level unlearning of Effusion for three co-occurring diseases: original image, Effusion attribution, retained-disease attribution, preserved shared tokens and suppressed Effusion tokens.

5 Conclusion

We studied selective disease unlearning in multi-label chest X-ray recognition, where co-occurring diseases share evidence, and showed that parameter-level and feature-level methods occupy the extremes of the forgetting–retention trade-off. MIST localises forgetting to target-specific ViT tokens and, unlike global optimisation-based methods, explicitly preserves the token representations that co-occurring diseases share. On NIH ChestX-ray14 it lowers Effusion AUROC from 0.870 to 0.403 with a retained drop of 0.007 and a co-occurrence drop of 0.008, 65% and 69% lower than SalUn at comparable forgetting, while modifying about 37% of the patch tokens, and the behaviour holds across five forget targets.

Ethics statement

This work uses the publicly released NIH ChestX-ray14 dataset (Wang et al., 2017), which contains de-identified chest radiographs; no new data were collected and no patient can be identified from the images or labels used. Selective unlearning supports the removal of a disease concept from a trained model in response to privacy requests, annotation errors or changes in clinical standards, and the method is intended for research use and not for clinical decision-making.

Reproducibility statement

Section 4 specifies the dataset, the patient-level five-fold splits, the backbone, the teacher training schedule, the optimiser and the loss settings, and Algorithm 1 gives the complete unlearning procedure. All results are reported as mean and standard deviation over the five folds.

AI use statement

In this work, we used generative AI tools to interpret results, by drafting and revising the text that reports them. We have not used generative AI tools to design the methodology or experiments or to implement methods, and generating synthetic data, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, translation, dataset cleaning and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit parts of the paper, to suggest its structure and to format references. We have reviewed all AI-assisted work: every number in the text, tables and figures was checked against the experiment outputs, and every figure against the method. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Ahmed Ashraf, Shehroz Khan, Nikhil Bhagwat, Mallar Chakravarty, and Babak Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *arXiv preprint arXiv:1812.01716*, 2018.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific Reports*, 9(1):6381, 2019.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and David Zhang. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2292–2302, 2020.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 896–911, 2021.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7766–7775, 2023.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023.

Zhipeng Deng, Luyang Luo, and Hao Chen. Enable the right to be forgotten with federated client unlearning in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 240–250. Springer, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Andreza Falcao and Filipe R Cordeiro. Does machine unlearning preserve clinical safety? a risk analysis for medical image classification. *arXiv preprint arXiv:2604.23854*, 2026.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12043–12051, 2024.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 792–801, 2021.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 130:259–266, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842. PMLR, 2020.

Misgina Tsighe Hagos, Kathleen M Curran, and Brian Mac Namee. Unlearning spurious correlations in chest x-ray classification. In *International Conference on Discovery Science*, pp. 387–397. Springer, 2023.

Shahad Hardan, Darya Taratynova, Abdelmajid Essofi, Karthik Nandakumar, and Mohammad Yaqub. Forget-MI: Machine unlearning for forgetting multimodal information in healthcare settings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 204–213. Springer, 2025.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2023.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36:1957–1987, 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Alessandro Mantelero. The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013.

Reza Nasirigerdeh, Nader Razmi, Julia A Schnabel, Daniel Rueckert, and Georgios Kaissis. Machine unlearning for medical imaging. *arXiv preprint arXiv:2407.07539*, 2024.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-024-00965-w.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. *Advances in Neural Information Processing Systems*, 34:13937–13949, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 145–158. Springer, 2011.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding factors influencing machine unlearning. In *IEEE European Symposium on Security and Privacy*, pp. 303–319, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Daniel Trippa, Cesare Campagnano, Maria Sofia Bucarelli, Gabriele Tolomei, and Fabrizio Silvestri. Gradient-based and task-agnostic machine unlearning. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *Network and Distributed System Security Symposium*, 2023.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023.

Yuyang Xue, Jingshuai Liu, Steven McDonagh, and Sotirios A Tsaftaris. Erase to enhance: Data-efficient machine unlearning in MRI reconstruction. *arXiv preprint arXiv:2405.15517*, 2024.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018.

A Training Procedure

Algorithm 1 summarises the complete unlearning procedure of MIST.

Algorithm 1 MIST: minimal intervention selective token unlearning

Require: frozen teacher F_T, target disease d_f, sets $\mathcal{D}_f, \mathcal{D}_r, \mathcal{D}_c$, thresholds ρ, τ, δ, weights $\lambda_{\text{conf}}, \lambda_{\text{Dist}}, \lambda_{\text{Shr}}$
Ensure: unlearned student F_S

$F_S \leftarrow F_T$ ▷ initialise the student from the teacher
for each epoch and each mini-batch drawn from $\mathcal{D}_f$, $\mathcal{D}_r$ and $\mathcal{D}_c$ **do**
  $H_f, H_r \leftarrow$ token attribution of F_T for d_f and for the retained diseases (Equation (3))
  $U \leftarrow H_f \odot (1 - H_r)$; $T_f \leftarrow \text{TopK}(U, \rho)$ ▷ forget-specific tokens
  $M_{\text{shared}} \leftarrow \mathbf{1}(H_f > \tau) \odot \mathbf{1}(H_r > \tau)$ ▷ shared tokens
  $\mathcal{L}_{\text{unlearn}} \leftarrow -\text{BCE}(F_S(x_f), y_f) + \lambda_{\text{conf}} \max(0, p_f - \delta)$ on T_f
  $\mathcal{L}_{\text{Dist}} \leftarrow \text{MSE}(F_T(I_r, T_r), F_S(I_r, T_r))$
  $\mathcal{L}_{\text{Shr}} \leftarrow \text{MSE}(M_{\text{shared}} \odot F_T(I_c, T_c),\ M_{\text{shared}} \odot F_S(I_c, T_c))$
  update F_S with $\nabla(\mathcal{L}_{\text{unlearn}} + \lambda_{\text{Dist}}\mathcal{L}_{\text{Dist}} + \lambda_{\text{Shr}}\mathcal{L}_{\text{Shr}})$
end for