

# Análise de Desaprendizado de Máquina em Modelos de Classificação de Imagens Médicas

Andreza M. C. Falcao<sup>1</sup>, Filipe R. Cordeiro<sup>2</sup>

<sup>1</sup> Visual Computing Lab, Departamento de Computação,  
Universidade Federal Rural de Pernambuco (UFRPE), Brasil

andreza.mcfalcao@urfpe.com, filipe.rolim@ufrpe.com

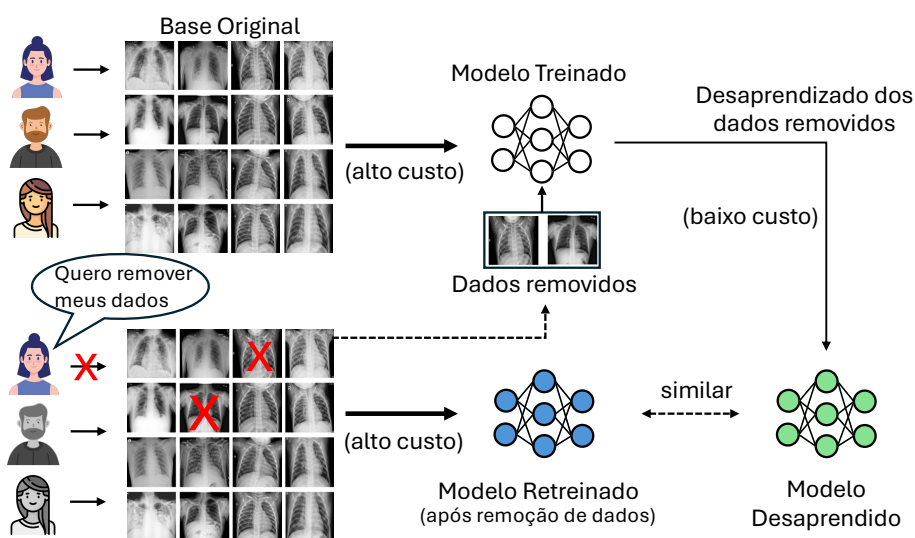
**Abstract.** *Machine unlearning aims to remove private or sensitive data from a pre-trained model while preserving the model's robustness. Despite recent advances, this technique has not been explored in medical image classification. This work evaluates the SalUn unlearning model by conducting experiments on the PathMNIST, OrganAMNIST, and BloodMNIST datasets. We also analyze the impact of data augmentation on the quality of unlearning. Results show that SalUn achieves performance close to full retraining, indicating an efficient solution for use in medical applications.*

**Resumo.** *O desaprendizado de máquina tem como objetivo remover dados privados ou sensíveis de um modelo pré-treinado, preservando a robustez do modelo. Apesar dos avanços, essa técnica não tem sido explorada em classificação de imagens médicas. Esse trabalho avalia o modelo de desaprendizagem Salun, conduzindo experimentos nas bases PathMNIST, OrganAMNIST e BloodMNIST. Também analisamos a influência do aumento de dados na qualidade do desaprendizado. Resultados mostram que o Salun obtém resultados próximos ao retreinamento completo, indicando uma solução eficiente para ser usada em aplicações médicas.*

## 1. Introdução

Modelos de aprendizagem de máquina têm sido amplamente utilizados na área médica para tarefas de detecção e auxílio ao diagnóstico a partir de imagens [Chan et al. 2020]. No entanto, esses modelos dependem de grandes volumes de dados para treinamento, os quais frequentemente contêm informações sensíveis de pacientes. Com o crescente uso de dados pessoais na medicina, surgem preocupações com a privacidade, impulsionadas por regulamentações como o Direito de Ser Esquecido (Right to be Forgotten)[Hoofnagle et al. 2019], que garantem aos usuários o direito de solicitar a remoção de seus dados pessoais de sistemas organizacionais[Dang 2021].

Tradicionalmente, a forma de lidar com essa questão envolve o retreinamento completo do modelo, excluindo os dados solicitados. Embora esse método garanta conformidade com as regulamentações, ele é extremamente custoso em termos de tempo de processamento e recursos computacionais. A área de Desaprendizado de Máquina (DM) investiga métodos para remover seletivamente a influência de dados específicos de modelos já treinados, sem a necessidade de retreinamento completo. A Figura 1 ilustra o processo de desaprendizado, onde, uma vez que um usuário solicita a remoção de um dado pessoal, o processo de desaprendizado de máquina busca remover a influência daqueles dados,



**Figura 1. Processo de desaprendizado de máquina. Ao ser solicitada a remoção de dados, é gerado um modelo desaprendido, sem a influência das amostras removidas. O desaprendizado tem como objetivo gerar um modelo próximo ao de um modelo retreinado do zero, mas com um baixo custo computacional.**

para que se comporte como um modelo retreinado sem os dados, mas sem o alto custo computacional.

Vários trabalhos têm sido propostos a fim de investigar a eficiência de métodos de DM em tarefas de classificação de propósito geral (objetos, animais, veículos, etc) [Zhang et al. 2023]. No entanto, esse tipo de técnica foi pouco explorado para tarefas de área médica e sabe-se que modelos de DM enfrentam desafios significativos ao lidar com problemas complexos.

Neste trabalho, investigamos a viabilidade do uso de métodos de DM como uma alternativa eficiente ao retreinamento completo na remoção de dados em modelos de classificação de imagens médicas. Para isso, nos baseamos no modelo Salun [Fan et al. 2024], que é um modelo de DM do estado da arte, aplicado para desaprender dados de um modelo de classificação ResNet18, utilizando as bases PathMNIST, OrganMNIST e BloodMNIST [Yang et al. 2021]. Além disso, investigamos a influência do uso de aumento de dados no desaprendizado, que também é algo investigado na literatura.

## 2. Trabalhos Relacionados

Trabalhos anteriores buscam reduzir a influência dos dados removidos através de mudança aleatória de anotação [Golatkhar et al. 2020] ou gradiente ascendente [Graves et al. 2021]. O modelo Saliency Unlearning (SalUn) [Fan et al. 2024] é um modelo de DM aproximado do estado da arte, que utiliza mapas de saliência para ajustar seletivamente pesos do modelo, baseado na ativação das amostras a serem removidas. Outros trabalhos buscam otimizar o processo de desaprendizado. Jia *et. al* [Jia et al. 2023] mostram que o processo de desaprendizado é mais eficiente ao realizar podas na rede, tornando a rede com menor número de parâmetros. O trabalho de Di *et. al* [Di et al. 2024] mostra que utilizar *soft labels* (anotações de entre 0 e 1) ajuda o modelo de desaprendizagem comparado ao

tradicional *hard label* (anotações com 0s e 1s). Neste trabalho, fazemos uma investigação também do impacto do uso de aumento de dados no processo de desaprendizado.

Apesar dos avanços, técnicas de DM ainda não foram avaliadas para bases de imagens médicas, que apresentam maior complexidade. Este trabalho busca preencher essa lacuna, em modelos treinados sobre MedMNIST, avaliando também a influência do uso de aumento de dados no processo de desaprendizagem.

### 3. Metodologia

#### 3.1. Base de dados

Utilizamos as bases de imagens PathMNIST, OrganAMNIST e BloodMNIST do repositório MedMNIST [Yang et al. 2021]. O MedMNIST é uma coleção de bases de dados de imagens médicas pré-processadas e em formato padronizado, projetada para facilitar a pesquisa em aprendizado de máquina em aplicações médicas.

A base PathMNIST contém imagens de biópsias de tecidos patológicos, coloridas, organizadas em 9 classes, que representam diferentes tipos de tecidos ou estados patológicos. A base contém 107.180 imagens, divididas em 89.996 imagens de treino, 10.004 de validação e 7.180 de teste.

A base OrganAMNIST contém imagens de seções axiais de ressonâncias magnéticas (MRI) de órgãos abdominais. A base é composta de 11 classes correspondendo a diferentes órgãos abdominais. A base é composta por 58.830 imagens, divididas em 34.561 imagens de treino, 6.491 de validação e 17.778 de teste.

A base BloodMNIST contém imagens microscópicas de células sanguíneas, com o objetivo de classificar diferentes tipos de células, auxiliando no diagnóstico de doenças hematológicas. A base possui 8 classes e 17.092 imagens, divididas em 11.959 imagens de treino, 1.712 de validação e 3.421 de teste.

Todas as imagens das bases foram redimensionadas para tamanho  $64 \times 64$  pixels.

#### 3.2. Treinamento e Desaprendizado

O modelo utilizado para a tarefa de classificação de imagens nas bases do MEDMNIST foi uma rede ResNet-18 [Wu et al. 2019]. Inicialmente, o modelo é treinado com a base inteira, por 200 épocas, utilizando taxa de aprendizagem 0.1, tamanho de *batch* 256. Utilizamos a configuração padrão de aumento de dados utilizada no Salun, que é o *random crop* e *horizontal flip* [Mumuni and Mumuni 2022]. Após as 200 épocas, o modelo treinado  $\theta_i$  é utilizado como base para a estratégia de desaprendizado.

A partir do modelo treinado com a base inteira  $\mathcal{D}_i$ , realizamos a operação de esquecimento de dados, que consiste em remover as amostras selecionadas de  $\mathcal{D}_i$ , resultando na base restante  $\mathcal{D}_r = \mathcal{D}_i - \mathcal{D}_f$ , onde  $\mathcal{D}_f$  consiste no conjunto de amostras a serem esquecidas. As amostras a serem esquecidas são selecionadas utilizando uma taxa de esquecimento  $\delta_f$ , realizando uma seleção de forma aleatória. Essa abordagem é a mesma utilizada em [Fan et al. 2024].

Utilizamos o método Salun para remover a influência das amostras em  $\mathcal{D}_f$  em  $\theta_i$ , obtendo o modelo desaprendido  $\theta_u = \text{Salun}(\mathcal{D}_f, \mathcal{D}_r, \theta_i)$ . O SalUn identifica as partes mais influentes do modelo para os dados a serem esquecidos e as modifica para reduzir essa

influência, permitindo que o modelo esqueça os dados indesejados de forma mais eficiente e com menor impacto em seu desempenho geral. Essa identificação é feita através da análise da saliência dos pesos do modelo em relação aos dados a serem esquecidos. O processo de desaprendizado é feito por 10 épocas, conforme utilizado em [Fan et al. 2024].

Também comparamos o Salun com o retreino completo. No retreinamento completo, é treinada uma nova ResNet18 por 200 épocas, utilizando a base  $\mathcal{D}_r$ , sem os dados esquecidos. Os resultados do retreino são usados como base (padrão ouro) para avaliar o modelo de desaprendizado.

### 3.3. Métricas de Avaliação

A métricas de avaliação de desaprendizado de máquina consistem em comparar os resultados do modelo desaprendido com o modelo retreinado. Nesse trabalho, utilizamos as métricas UA, RA, TA, MIA, AG e RTE, as quais são utilizadas em [Fan et al. 2024, Di et al. 2024]. Cada métrica é descrita a seguir:

- **Unlearning Accuracy (UA):** Mede a acurácia do modelo desaprendido nos dados a serem esquecidos. Um valor baixo de UA indica que o modelo conseguiu esquecer os dados indesejados.
- **Remaining Accuracy (RA):** Mede a acurácia do modelo desaprendido nos dados restantes (os dados que não foram esquecidos). Um valor alto de RA indica que o modelo conseguiu manter seu desempenho nos dados relevantes.
- **Testing Accuracy (TA):** Mede a acurácia do modelo desaprendido em um conjunto de teste independente. Um valor alto de TA indica que o modelo conseguiu generalizar para novos dados.
- **Membership Inference Attack (MIA):** Mede a vulnerabilidade do modelo desaprendido a ataques de inferência de associação. Um valor baixo de MIA indica que o modelo está menos vulnerável a esses ataques.
- **Average GAP (AG):** Mede a proximidade do modelo desaprendido com o modelo retreinado. É calculado pelo módulo da média das diferenças entre as métricas UA, RA, TA e MU do modelo desaprendido e o modelo retreinado.
- **Run-time efficiency (RTE):** Mede o tempo de execução de cada método.

## 4. Resultados

Analizamos o desempenho do Salun para as bases BloodMNIST, OrganAMNIST e PathMNIST, considerando uma taxa de esquecimento  $\delta$  dos dados de treino de 10% e 50%. Os resultados para  $\delta = 10\%$  e  $\delta = 50\%$  são mostrados nas Tabelas 1 e 2, respectivamente. Os valores em negrito mostram a diferença comparada com o método de Retreino. Os resultados mostram uma diferença média próxima de zero, representada pela métrica AG, o que indica uma proximidade de resultados do método de retreino. No entanto, o tempo de execução é muito menor utilizando o Salun, como mostra a métrica RTE. Os resultados também indicam uma maior dificuldade de desaprendizado para uma taxa de esquecimento maior e também para a base PathMNIST, que é mais complexa, como mostram os resultados da métrica TA.

Analizamos também o impacto de usar *data augmentation* durante as etapas de treinamento e desaprendizagem. Para isso, analisamos 3 cenários de uso de *data augmentation*:

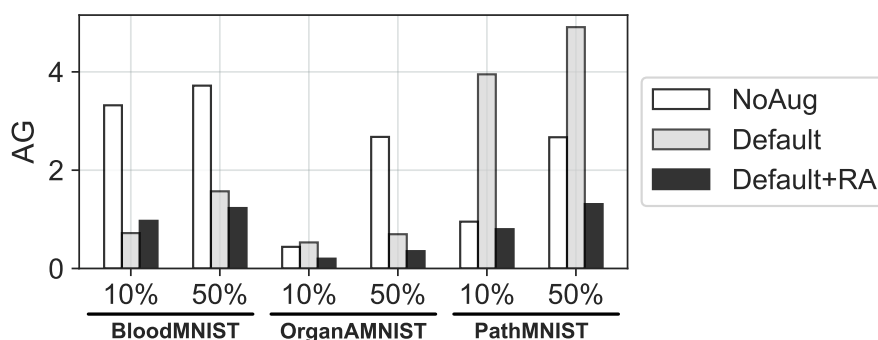
**Tabela 1. Resultados para taxa de esquecimento de 10%, para as métricas UA, RA, TA, MIA, AG e RTE (em minutos). Resultados em negrito mostram a diferença de cada métrica comparado com o método de Retreino, que é o padrão ouro. A métrica AG mostra a média dessas diferenças.**

Base	Método	Taxa de Esquecimento (10%)					
		UA	RA	TA	MIA	AG	RTE
BloodMNIST	Retreino	0,84 <b>(0,00)</b>	99,80 <b>(0,00)</b>	98,57 <b>(0,00)</b>	1,76 <b>(0,00)</b>	<b>0,00</b>	22,2
	Salun	0,00 <b>(0,84)</b>	99,92 <b>(0,12)</b>	98,89 <b>(0,32)</b>	0,17 <b>(1,59)</b>	<b>0,72</b>	1,1
OrganAMNIST	Retreino	0,06 <b>(0,00)</b>	100,00 <b>(0,00)</b>	96,37 <b>(0,00)</b>	1,53 <b>(0,00)</b>	<b>0,00</b>	63,3
	Salun	0,00 <b>(0,06)</b>	100,00 <b>(0,00)</b>	95,13 <b>(1,24)</b>	0,69 <b>(0,84)</b>	<b>0,53</b>	2,8
PathMNIST	Retreino	0,11 <b>(0,00)</b>	100,00 <b>(0,00)</b>	87,77 <b>(0,00)</b>	1,06 <b>(0,00)</b>	<b>0,00</b>	160
	Salun	1,09 <b>(0,98)</b>	98,84 <b>(1,16)</b>	77,49 <b>(10,28)</b>	4,43 <b>(3,37)</b>	<b>3,95</b>	7,6

**Tabela 2. Resultados para taxa de esquecimento de 50%, para as métricas UA, RA, TA, MIA, AG e RTE (em minutos). Resultados em negrito mostram a diferença de cada métrica comparado com o método de Retreino, que é o padrão ouro. A métrica AG mostra a média dessas diferenças.**

Base	Método	Taxa de Esquecimento (50%)					
		UA	RA	TA	MIA	AG	RTE
BloodMNIST	Retreino	1,37 <b>(0,00)</b>	100,00 <b>(0,00)</b>	98,48 <b>(0,00)</b>	3,61 <b>(0,00)</b>	<b>0,00</b>	22,2
	Salun	0,12 <b>(1,25)</b>	99,93 <b>(0,07)</b>	98,77 <b>(0,29)</b>	0,45 <b>(3,16)</b>	<b>1,57</b>	1,3
OrganAMNIST	Retreino	0,13 <b>(0,00)</b>	100,00 <b>(0,00)</b>	95,93 <b>(0,00)</b>	1,93 <b>(0,00)</b>	<b>0,00</b>	63,3
	Salun	0,02 <b>(0,11)</b>	99,94 <b>(0,06)</b>	94,99 <b>(0,94)</b>	0,89 <b>(1,04)</b>	<b>0,70</b>	3,5
PathMNIST	Retreino	0,20 <b>(0,00)</b>	100,00 <b>(0,00)</b>	91,80 <b>(0,00)</b>	1,93 <b>(0,00)</b>	<b>0,00</b>	160
	Salun	2,33 <b>(2,13)</b>	97,82 <b>(2,18)</b>	83,87 <b>(7,93)</b>	6,60 <b>(4,67)</b>	<b>4,91</b>	8,9

1) *NoAug* (sem augmentation), 2) *Default*: *random crop* + horizontal flip, 3) e 3) *Default* + RA: *random crop* + horizontal flip + RandomAug [Mumuni and Mumuni 2022]. A configuração *default* é a configuração do Salun e a usada na maioria dos trabalhos da literatura. A combinação com RandomAug mostrou melhoria para a maioria das análises, conforme mostra a Figura 2.



**Figura 2. Resultados do método Salun, observando a métrica AG, usando 3 cenários de aumento de dados: NoAug, Default e Default+RA, considerando níveis de esquecimento de 10% e 50%.**

## 5. Conclusão

O método SalUn obteve um desempenho comparável ao do Retrain nas bases analisadas: PathMNIST, OrganMNIST e BloodMNIST. As métricas analisadas indicam que o SalUn é eficaz em remover a influência de dados específicos sem comprometer significativamente o desempenho geral do modelo. Também concluímos que o uso de aumento de dados pode auxiliar na eficiência do desaprendizado de máquina e que esse modelo pode ser usado com eficiência em bases de dados médicas.

## Referências

- Chan, H.-P., Samala, R. K., Hadjiiski, L. M., and Zhou, C. (2020). Deep learning in medical image analysis. *Deep learning in medical image analysis: challenges and applications*, pages 3–21.
- Dang, Q.-V. (2021). Right to be forgotten in the age of machine learning. In *Advances in Digital Science: ICADS 2021*, pages 403–411. Springer.
- Di, Z., Zhu, Z., Jia, J., Liu, J., Takhirov, Z., Jiang, B., Yao, Y., Liu, S., and Liu, Y. (2024). Label smoothing improves machine unlearning. *arXiv preprint arXiv:2406.07698*.
- Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. (2024). Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*.
- Golatkhar, A., Achille, A., and Soatto, S. (2020). Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312.
- Graves, L., Nagisetty, V., and Ganesh, V. (2021). Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524.
- Hoofnagle, C. J., Van Der Sloot, B., and Borgesius, F. Z. (2019). The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.
- Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., and Liu, S. (2023). Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605.
- Mumuni, A. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.
- Wu, Z., Shen, C., and Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern recognition*, 90:119–133.
- Yang, J., Shi, R., Wei, D., Liu, Z., Wang, L., Zhou, Y., Zhou, S., Bian, C., Li, L., Wang, X., et al. (2021). Medmnist: A lightweight automl benchmark for medical image analysis. <https://medmnist.com>. Accessed: February 13, 2025.
- Zhang, H., Nakamura, T., Isohara, T., and Sakurai, K. (2023). A review on machine unlearning. *SN Computer Science*, 4(4):337.