Why patient data cannot be easily forgotten?

Ruolin Su[*1] R.SU-1@SMS.ED.AC.UK
Xiao Liu[*1] XIAO.LIU@ED.AC.UK
Sotirios A. Tsaftaris[2] S.TSAFTARIS@ED.AC.UK
[1] *School of Engineering, University of Edinburgh, West Mains Rd, Edinburgh EH9 3FB, UK*
[2] *The Alan Turing Institute, London, UK*

Editors: Under Review for MIDL 2022

Abstract

Rights provisioned within data protection regulations, permit patients to request that knowledge about their information be eliminated by data holders. With the advent of AI learned on data, one can imagine that such rights can extent to requests for forgetting knowledge of patient's data within AI models. This is motivated by recent findings that AI models can memorise information about data and have been shown to be vulnerable to methods that can aim to uncover specific and private information from a model. Forgetting patients' imaging data from AI models, is still an under-explored problem. In this paper, we study the influence of patient data on model performance and formulate two hypotheses for a patient's data: either they are common and similar to other patients or form edge cases, i.e. unique and rare cases. This shows that it is not possible to easily *forget patient data*. We propose a targeted forgetting approach to perform patient-wise forgetting. Extensive experiments on the benchmark Automated Cardiac Diagnosis Challenge (ACDC) dataset, a medical dataset composed of cardiac MRI images, showcase the improved forgetting performance of the proposed targeted forgetting approach as opposed to a state-of-the-art method.

Keywords: Privacy, Patient-wise Forgetting, Scrubbing, Learning

1. Introduction

Apart from solely improving algorithm performance, developing trusted deep learning algorithms that respect data privacy has now become of crucial importance (Abadi et al., 2016; Liu and Tsaftaris, 2020). It is now known that deep models can memorise a user's sensitive information (Arpit et al., 2017). Several attack types (Truex et al., 2019) including simple reverse engineering (Fredrikson et al., 2015) can reveal private information. For example, using the attacks discussed in (Wu et al., 2020), by looking at the output and weights of a model, sensitive patient information such as gender, age, BMI etc. can be extracted. It is then without surprise why a user (e.g. a patient) may require that private information is not only deleted from databases but that any such information is forgotten by AI models trained on such databases.

A naive solution to forget a patient's information is to re-train the model without the concerning data. However, re-training is extremely time-consuming and sometimes impossible (Shintre et al., 2019). For example, in a federated learning scheme (McMahan et al.,

[*] Contributed equally

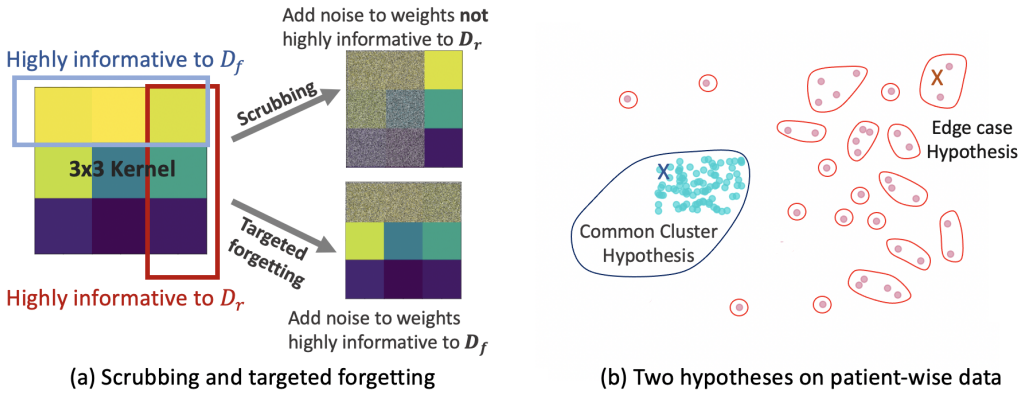


Figure 1: (a) Visualisation of the scrubbing and targeted forgetting methods. $\mathcal{D}_r$ and $\mathcal{D}_f$ are the retaining data and the forgetting data. (b) Illustration of the two hypotheses. Blue contour delineates a big sub-population of similar samples within a *common cluster*; red contours denote several small sub-populations of distinct samples in *edge cases*. X and X are examples of samples to be forgotten.

2017), the data are not centrally aggregated but retained in servers (e.g. distributed in different hospitals) which may not be available anymore to participate in re-training.

Machine unlearning/forgetting aims to remove private information of concerning data without re-training the model. This involves post-processing to the trained model to make it act like a re-trained one that has never seen the concerning data. Several studies have previously explored forgetting/unlearning/deleting data and made remarkable progress (Cao and Yang, 2015; Ginart et al., 2019; Golatkar et al., 2020; Nguyen et al., 2020; Sekhari et al., 2021). When the concept of machine unlearning/forgetting was first developed by Cao and Yang (2015), they discussed forgetting in statistical query learning (Kearns, 1998). Ginart et al. (2019) specifically deal with data deletion in k-means clustering with excellent deleting efficiency. Another approach is to rely on variational inference and Bayesian models (Nguyen et al., 2020). Recently, Sekhari et al. (2021) provide a data deleting algorithm by expanding the forgetting limit whilst considering the model's generalization ability. Golatkar et al. (2020) address machine unlearning on deep networks to forget a subset of training data (e.g. one class of CIFAR-10 (Krizhevsky et al., 2009)). Their scrubbing approach, shown in Fig. 1(a), adds weighted noise to model weights which are uninformative to the remaining data (training data excluding the concerning data) to achieve a weaker form of differential privacy (Dwork et al., 2014).

Patient-wise forgetting Different from previous works, here we specifically consider the scenario of patient-wise forgetting, where instead of forgetting a selected random cohort of data, the data to be forgotten originate from a single patient. To address this problem, we take inspiration by the scrubbing procedure of Golatkar et al. (2020). However, our experiments illustrate that performance varies between different patients: from adequate forgetting performance to unacceptable performance (either in forgetting or generalisation to test data). We show that this is due to assumptions made by Golatkar et al. (2020) on how

data are distributed. Hence, simple translation of existing machine unlearning/forgetting methods to patient-wise forgetting is not straightforward.

A tale of two hypotheses We hypothesise (and show experimentally) that in a medical dataset, a patient's data can either be similar to other data (and form clusters) or form edge cases as we depict in Fig. 1(b). These hypotheses are aligned with recent studies on long-tail learning (Buda et al., 2018; Liu et al., 2019), where different sub-populations within a class can exist with some being in the so-called long tail.[1] Subsequently, we will refer to these cases as *common cluster* and *edge case* hypotheses. For patients under the two different hypotheses, forgetting and generalisation performance obtained after scrubbing (Golatkar et al., 2020) vary as detailed in Section 3. For the scrubbing method, it removes information not highly related to most of the remaining data to maintain good generalisation after forgetting. When forgetting a patient under common cluster hypothesis, scrubbing has to sacrifice model generalisation to forget this patient. Because by forgetting this patient's data will necessitate that a whole cluster in the remaining data should also be forgotten. When forgetting an edge-case patient, the scrubbing method does not remove specifically the edge-case patient's information but noise will be introduced to model weights corresponding to most of the edge cases in the remaining dataset. Hence, the overall model performance will be negatively affected.

The proposed targeted forgetting To achieve patient-wise forgetting, we propose targeted forgetting, which only adds weighted noise to highly informative (i.e. targeted) model weights of a forgetting patient. Here, we follow Golatkar et al. (2020) to measure the informativeness of model weights with Fisher Information Matrix (FIM), which determines the strength of noise to be added to different model weights. With our proposed targeted forgetting, we can precisely forget data of patients within edge case hypothesis and maintain good model generalisation performance. For data of patients falling under the common cluster hypothesis, the algorithm can forget the corresponding information though the model performance on the whole cluster will be affected.[2] **Contributions**:

1. We introduce the problem of patient-wise forgetting.
2. We observe variable performance in patient-wise forgetting with the method in Golatkar et al. (2020) and formulate two hypotheses that explain this behaviour.
3. We propose a new targeted forgetting method and perform extensive experiments on a medical benchmark dataset to showcase improved patient-wise forgetting performance.

Our work we hope will inspire future research to consider how different data affect forgetting methods especially in a patient-wise forgetting setting.

2. Method

We adopt the notation of Golatkar et al. (2020). Given a *training* dataset $\mathcal{D}$, a *forgetting* subset $\mathcal{D}_f \subset \mathcal{D}$ contains the images we desire to be removed from a model trained on $\mathcal{D}$.

1. There is a connection between edge cases and active learning (Settles, 2009), where one aims to actively label diverse data to bring more information to the model and achieve better accuracy.
2. This implies that for some patients especially within the common cluster hypothesis, it is not easy (probably impossible) to forget the corresponding information without negatively affecting the model.

The *retaining* dataset is the complement $\mathcal{D}_r$=$\mathcal{D} \setminus \mathcal{D}_f$, and thus $\mathcal{D}_r \cap \mathcal{D}_f = \emptyset$. We denote the test data as $\mathcal{D}_{test}$. For patient-wise forgetting, we set $\mathcal{D}_f$ to be all the images of one patient. Let $\mathbf{w}$ be the weights of a model $A(\mathcal{D})$, which is trained on $\mathcal{D}$ using any stochastic learning algorithm $A(\cdot)$. Let $S(\mathbf{w})$ denote the operations applied to model weights to forget $\mathcal{D}_f$ in the model. We define the *golden standard* model as $A(\mathcal{D}_r)$.

2.1. The scrubbing method

Assuming that $A(\mathcal{D})$ and $\mathcal{D}_r$ are accessible, Golatkar et al. (2020) propose a robust scrubbing procedure modifying model $A(\mathcal{D})$, to brings the model closer to a golden standard model $A(\mathcal{D}_r)$. They use the Fisher Information Matrix (FIM) to approximate the hessian of the loss on $\mathcal{D}_r$, where higher values in FIM (see Appendix B for the calculation) denote higher correlation between corresponding model weights and $\mathcal{D}_r$. With the FIM, they introduce different noise strength to model weights to remove information not highly informative to $\mathcal{D}_r$, and thus forget information corresponding to $\mathcal{D}_f$. The scrubbing function is defined as:

$$\mathrm{S}(\mathbf{w}) = \mathbf{w} + \left(\lambda\sigma_h^2\right)^{\frac{1}{4}} F_{\mathcal{D}_r}(\mathbf{w})^{-1/4}, \tag{1}$$

where $F_{\mathcal{D}_r}(\mathbf{w})$ denotes the FIM computed for $\mathbf{w}$ on $\mathcal{D}_r$. Scrubbing is controlled by two hyperparameters: λ decides the scale of noise introduced to $\mathbf{w}$ therefore it controls the model accuracy on $\mathcal{D}_f$; σ_h is a normal distributed error term which simulates the error of the stochastic algorithm, ensuring a continuous gradient flow after the scrubbing procedure. Practically during experiments, the product of the two hyperparameters is tuned as a whole.
What happens during patient-wise forgetting? The FIM ($F_{\mathcal{D}_r}(\mathbf{w})$) for a medical dataset is derived by summing up the FIM of each patient's data. Weights highly related to the cluster's features show high values in FIM because several cluster patients within $\mathcal{D}_r$ are correlated to these weights. Whereas for edge cases, no other patients are correlated with the same weights as of these edge cases; thus, the aggregated values in FIM for these weights corresponding to all edge cases are relatively small. When forgetting a common cluster case, scrubbing by trying to preserve model performance on $\mathcal{D}_r$ sacrifices forgetting performance. When forgetting an edge case, the model performance on edge cases in $\mathcal{D}_r$ will be negatively affected.

2.2. The targeted forgetting method

Based on the idea of scrubbing model weights, and the connection between the hessian of a loss on a set of data of a model and the extent to which the weights are informative about these data, we develop the targeted forgetting procedure. We assume access to the forgetting data $\mathcal{D}_f$ instead of $\mathcal{D}_r$. We believe that even in a real patient-wise forgetting scenario, temporary access to patient data is permissible until forgetting is achieved.

We compute FIM for $\mathbf{w}$ on $\mathcal{D}_f$ instead of $\mathcal{D}_r$ to approximate the noise added to model weights. Instead of keeping the most informative model weights corresponding to $\mathcal{D}_r$ as in Golatkar et al. (2020), our targeted forgetting can be considered as precisely introducing noise to model weights highly informative about $\mathcal{D}_f$ (see Fig. 1(a)). Our proposed targeted forgetting is defined as:

$$\mathrm{S_T}(\mathbf{w}) = \mathbf{w} + \left(\lambda_T\sigma_{h_T}^2\right)^{\frac{1}{4}} F_{\mathcal{D}_f}(\mathbf{w})^{1/4}, \tag{2}$$

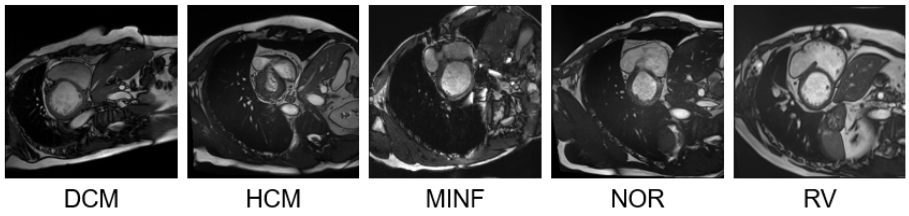


Figure 2: Example images of ACDC dataset. DCM: dilated cardiomyopathy. HCM: hypertrophic cardiomyopathy. MINF: myocardial infarction. NOR: normal subjects. RV: abnormal right ventricle.

where λ_T and σ_{h_T} are analogous parameters to those defined as λ and σ_h in Eq 1.

Performance on the two hypotheses We briefly discuss the expectations of targeted forgetting performance here in anticipation of the detailed results and discussion of Section 3. *Common cluster hypothesis*: Targeted forgetting will add noise to the most informative model weights corresponding to $\mathcal{D}_f$ so it will reduce model performance on the corresponding cluster in $\mathcal{D}_r$ and $\mathcal{D}_{test}$. *Edge case hypothesis*: Targeted forgetting will precisely remove information of one edge case and maintain good model performance.

3. Experiments

We first explore why the scrubbing method works well on computer vision datasets but shows poorer performance on patient-wise medical data forgetting. We conduct an experiment to demonstrate the intrinsic dataset biases of CIFAR-10 and ACDC (Bernard et al., 2018). Then, we compare the forgetting and model performance after forgetting achieved using the scrubbing and our targeted forgetting methods.

Datasets: CIFAR-10 has 60,000 images (size 32×32) of 10-class objects. The Automated Cardiac Diagnosis Challenge (ACDC) dataset contains 4D cardiac data from 100 patients with four pathologies classes and a normal group. We split the 100 patients into training and testing subsets. Overall, by preprocess the patient data into 224x224 2D images, there are 14,724 images from 90 patients in the training set D, and 1,464 images from 10 patients in the testing set $\mathcal{D}_{test}$. Patients in both sets are equally distributed across the five classes. Example images from the ACDC dataset are shown in Fig. 2. When conducting experiments under the patient-wise forgetting scenario, we only select one patient to be forgotten devising the forgetting set composed of all the images of the same patient.

Implementation details: For CIFAR-10, we follow the implementation steps in Golatkar et al. (2020). When training the ACDC classifier, the model has a VGG-like architecture as in Thermos et al. (2021). We use Cross Entropy as the loss function and use Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.5$, $\beta_2 = 0.999$. During training we use data augmentation including random rotation, Gaussian blur, horizontal and vertical flip. We train all classifiers with a learning rate of 0.0001 for 13 epochs. The original model trained with all 90 patients has 0.00 error on $\mathcal{D}_r$ and $\mathcal{D}_f$, and 0.19 error on $\mathcal{D}_{test}$.

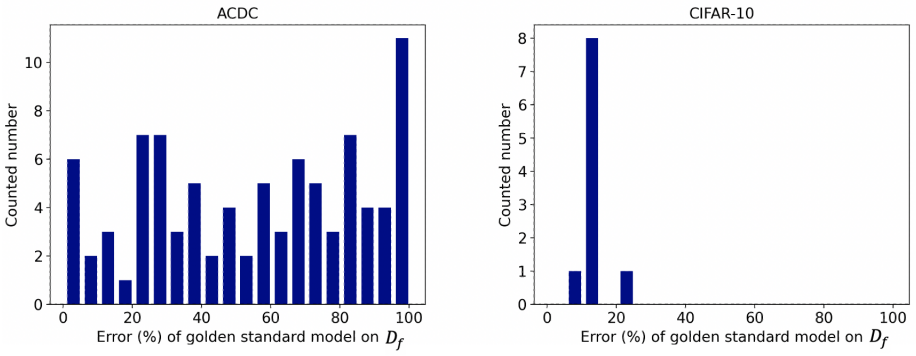


Figure 3: Histograms of re-training experiments. The y-axis refers to the total number of patients/sets whose golden standard lies within an interval(i.e. [95,100]) of x-axis.

3.1. The hardness of patient-wise forgetting

Here we want to show that some patient data are hard to learn and consequently make patient-wise forgetting hard. We compare between CIFAR-10 and ACDC. Given 90 patients in the training set of ACDC, we remove one patient's data and fully re-train a model on the remaining 89 i.e. the golden standard model. We then measure the golden model error on the deleted patient. We repeat this for all 90 patients. For CIFAR-10, we select 10 non-overlapping sets from its training set, each with 100 images from the same class, to be the deleting candidates and repeat the same process as in ACDC. If data are hard, error of the golden model will be high and thus should be harder to forget

Results and discussion: Fig. 3 collects the findings of this experiment as histograms and shows the differences between the two datasets. Overall, for ACDC, the 90 individually measured results of golden standard error on $\mathcal{D}_f$ vary from 0% to 100%, whereas in CIFAR-10, the 10 experimental results only vary from 10% to 25%. High golden model error on a $\mathcal{D}_f$ means that the model is unable to generalise to this patient's data; thus, this patient is not similar to any other patients in the training set, and must belong to the edge case hypothesis. By considering a threshold of 50% on the error of the golden model, we find that $> 60\%$ **of patients in ACDC can be considered to belong to the edge case hypothesis**. Appendix A (Fig. 4) showcases some examples. This is markedly different in CIFAR-10: golden model results concentrate at low error indicating that few edge cases exist. In addition, as we discussed in the introduction, scrubbing multiple weights that are informative about different edge cases in $\mathcal{D}_r$ will degrade model generalisation on both $\mathcal{D}_r$ and $\mathcal{D}_{test}$. This will explain the results of the scrubbing method: under-performance in ACDC because many patients fall under the edge case hypothesis.

3.2. Forgetting performance with respect to golden standard models

We assess forgetting performance by comparing against golden standard models. A method has good forgetting performance by coming as close to the performance of the golden standard model on $\mathcal{D}_f$. For simplicity we focus on four representative patients using the analysis in Section 3.1 (for more patients' results see Appendix C): patients 94 and 5 that fall under

Table 1: Forgetting results for four patients. We report Error = 1−Accuracy on the retaining ($\mathcal{D}_r$), forgetting ($\mathcal{D}_f$) and test ($\mathcal{D}_{test}$) sets respectively. Scrubbing Method refers to the method of Golatkar et al. (2020) whereas Targeted Forgetting refers to the method in Section 2.2. Red and blue denote upper and lower bounds for each row respectively, with performance being better when it is closer to the bounds.

$\mathcal{D}_f$ Patient ID	Error on	Golden Standard	Scrubbing Method	Targeted Forgetting
94	$\mathcal{D}_r$	0.00	0.46	0.23
	$\mathcal{D}_f$	1.00	1.00	1.00
	$\mathcal{D}_{test}$	0.24	0.56	0.30
5	$\mathcal{D}_r$	0.00	0.35	0.24
	$\mathcal{D}_f$	0.81	0.82	0.83
	$\mathcal{D}_{test}$	0.25	0.46	0.38
13	$\mathcal{D}_r$	0.00	0.08	0.21
	$\mathcal{D}_f$	0.20	0.20	0.20
	$\mathcal{D}_{test}$	0.19	0.30	0.37
9	$\mathcal{D}_r$	0.00	0.01	0.06
	$\mathcal{D}_f$	0.01	0.01	0.01
	$\mathcal{D}_{test}$	0.23	0.23	0.32

the edge case hypothesis; and patients 13 and 9, that fall under a common cluster hypothesis. We adjust the hyperparameters of both methods to make them achieve the same error on $\mathcal{D}_f$ as the golden standard. Then we compare model error on $\mathcal{D}_r$ and $\mathcal{D}_{test}$.

Results and discussion: Table 1 shows the results of forgetting $\mathcal{D}_f$. For patient 94, the golden standard has an error of 1.00. Both methods match this error. However, the scrubbing method's error on $\mathcal{D}_r$ (0.46) is very high indicating that to achieve such forgetting, training performance suffers. Meanwhile, generalisation performance also degrades (0.56 vs 0.24 of the golden standard). Targeted forgetting achieves good forgetting (1.00 error on $\mathcal{D}_f$), and good generalisation performance (0.30 on $\mathcal{D}_{test}$). Similar observations hold also for another edge case such as patient 5. For patients 13 and 9, the golden standard errors are relatively very low, 0.20 and 0.01. Scrubbing overall matches the forgetting performance of targeted forgetting. Targeted forgetting though shows worse generalisation performance.

3.3. Patient-wise forgetting performance without golden standard model

Here we consider a more stringent scenario: the re-trained golden standard model is not available for deciding level of noise to be added during scrubbing or forgetting. We adjust noise strength (low, medium and high) by modulating the hyperparameter λ in both methods to achieve different levels of forgetting. Additional implementation details are in the Appendix D. We show the results of the same four representative patients in Table 2.[3]

Results and discussion: The two forgetting methods vary in performance for patients under the different hypotheses. With targeted forgetting, we show better model generalisation on $\mathcal{D}_{test}$ for patient 94 at all noise levels compared with the scrubbing method (0.22

3. Note that some values within Table 2 are different from Table 1. This is because without any knowledge of forgetting, we have to add noise to a fixed number of weights. When obtaining Table 1 this is adjustable: prior knowledge of how much to forget is available via the golden standard.

Table 2: Forgetting results for four patients. We report Error = 1−Accuracy on the retaining ($\mathcal{D}_r$), forgetting ($\mathcal{D}_f$) and test ($\mathcal{D}_{test}$) sets respectively. With respect to error on $\mathcal{D}_f$ **High** noise level refers to the noise strength when a method reaches 1.00 error; **Medium**: 0.85±0.05 error; and **Low**: 0.14±0.05 error.

Patient ID	Error on	Golden Standard	Noise level					
			Low		Medium		High	
			Scrubbing	Targeted Forgetting	Scrubbing	Targeted Forgetting	Scrubbing	Targeted Forgetting
94	$\mathcal{D}_r$	0.00	0.29	0.03	0.60	0.14	0.76	0.26
	$\mathcal{D}_f$	1.00	0.15	0.17	0.86	0.83	1.00	1.00
	$\mathcal{D}_{test}$	0.24	0.67	0.22	0.74	0.29	0.75	0.31
5	$\mathcal{D}_r$	0.00	0.04	0.04	0.25	0.27	0.40	0.44
	$\mathcal{D}_f$	0.81	0.13	0.11	0.85	0.86	1.00	1.00
	$\mathcal{D}_{test}$	0.25	0.39	0.27	0.62	0.41	0.70	0.51
13	$\mathcal{D}_r$	0.00	0.02	0.15	0.21	0.48	0.36	0.61
	$\mathcal{D}_f$	0.20	0.11	0.09	0.87	0.85	1.00	1.00
	$\mathcal{D}_{test}$	0.19	0.36	0.34	0.59	0.52	0.69	0.60
9	$\mathcal{D}_r$	0.00	0.04	0.25	0.25	0.60	0.40	0.68
	$\mathcal{D}_f$	0.01	0.18	0.15	0.89	0.86	1.00	1.00
	$\mathcal{D}_{test}$	0.23	0.40	0.44	0.64	0.61	0.70	0.66

vs 0.67 at low, 0.29 vs 0.74 at medium, 0.31 vs 0.75 at high). Similar results hold also for patient 5. This showcases the ability of targeted forgetting in maintaining model generalisation when forgetting edge cases. For patient 9 (a common cluster case), the model error on $\mathcal{D}_r$ is better with the scrubbing method at all levels than with the targeted forgetting (low: 0.04 vs 0.25, medium: 0.25 vs 0.60, and high: 0.40 vs 0.68). However, the model's generalisation performance on $\mathcal{D}_{test}$ with the two methods, which is with more concern, has been very close all the time, with at low (0.40 and 0.44), medium (0.64 and 0.61) and high (0.70 and 0.66). Moreover, the error becomes relatively high especially at medium and high noise levels. Investigating the model performance on another common cluster patient 13, we also observe unsatisfactory performance on $\mathcal{D}_{test}$ at medium and high noise level, meaning that both methods sacrifice the model's generalisation to forget more about a patient under common cluster hypothesis, when not knowing how much to forget from a golden standard.

4. Conclusion

We consider patient-wise forgetting in deep learning models. Our experiments reveal that forgetting a patient's medical image data is harder than other vision domains. We found that this is due to data falling on two hypotheses: common cluster and edge case. We identified limitations of an existing state-of-the-art scrubbing method and proposed a new targeted forgetting approach. Experiments highlight the different roles of these two hypotheses and the importance of considering the dataset bias. We perform experiments on cardiac MRI data but our approach is data-agnostic, which we plan to apply on different medical datasets in the future. In addition, future research on patient-wise forgetting should focus on better ways of detecting which hypothesis the data of patients belong to and how to measure patient-wise forgetting performance with considering the two hypotheses.

Acknowledgments

This work was supported by the University of Edinburgh, the Royal Academy of Engineering and Canon Medical Research Europe by a PhD studentship to Xiao Liu. This work was partially supported by the Alan Turing Institute under EPSRC grant EP/N510129/1. S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\8\25) and the [in part] support of the Industrial Centre for AI Research in digitalDiagnostics (iCAIRD, https://icaird.com) which is funded by Innovate UK on behalf of UK Research and Innovation (UKRI) [project number: 104690].

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. doi: 10.1109/SP.2015.35.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Antonio Ginart, Melody Y Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning. *arXiv preprint arXiv:1907.05012*, 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*, pages 9304–9312, 2020.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *In proc. ICLR*, 2015.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Xiao Liu and Sotirios A Tsaftaris. Have you forgotten? a method to assess if machine learning models have forgotten data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2020.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33, 2020.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *arXiv preprint arXiv:2103.03279*, 2021.

Burr Settles. Active learning literature survey. 2009.

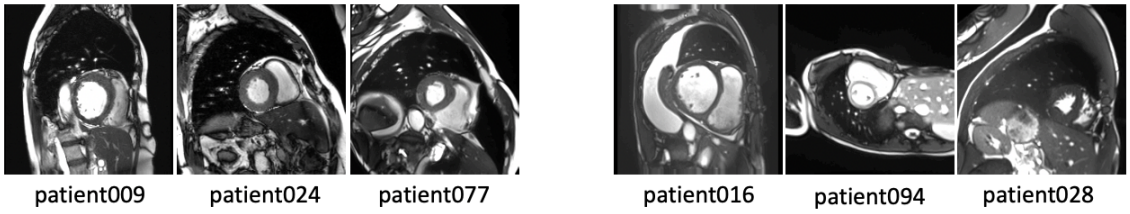
Saurabh Shintre, Kevin A Roundy, and Jasjeet Dhaliwal. Making machine learning forget. In *Annual Privacy Forum*, pages 72–83. Springer, 2019.

Spyridon Thermos, Xiao Liu, Alison O'Neil, and Sotirios A. Tsaftaris. Controllable cardiac synthesis via disentangled anatomy arithmetic. In *MICCAI*, 2021.

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019.

Maoqiang Wu, Xinyue Zhang, Jiahao Ding, Hien Nguyen, Rong Yu, Miao Pan, and Stephen T Wong. Evaluation of inference attack models for deep learning on medical data. *arXiv preprint arXiv:2011.00177*, 2020.

Appendix A. Visual examples of images in two hypotheses



(a) Common Cluster Hypothesis (b) Edge case Hypothesis

Figure 4: Visualisation of images in the ACDC dataset (a) Three patients fall under common cluster hypothesis (b) Three edge case examples.

Appendix B. Calculation of the Fisher Information Matrix

The Fisher Information Matrix F of a distribution $P_{x,y}(\mathbf{w})$ w.r.t. $\mathbf{w}$ defined in Golatkar et al. (2020) is:

$$F = \mathbb{E}_{x\sim\mathcal{D},y\sim p(y|x)}\left[\nabla_{\mathbf{w}} \log p_{\mathbf{w}}(y \mid x) \nabla_{\mathbf{w}} \log p_{\mathbf{w}}(y \mid x)^T\right] \tag{3}$$

During implementation, to save computational memory, only the diagonal values for FIM are computed and stored. The trace of FIM is calculated by taking the expectation of the outer product of the gradient of a deep learning model. Note that in our experiment, since we treat the medical dataset at patient-level, we take an extra normalisation step for each patient and take the expectation at patient-level, instead of taking the expectation over all data.

Because the FIM is computed to approximate the hessian matrix of model loss on a set $\mathcal{D}$, a value within FIM reflects to what extent the change to its corresponding weights would influence the model's classification process on this set. Hence, if a model weight is correlated with multiple data and thus considered to be important in classifying these data, its corresponding value in FIM would be relatively high, and vice versa.

Appendix C. More results for Table 1

Here we report more results with the same experimental settings in Section 3.2. The experiments include 10 patients with different golden standard errors in total. Four typical patients are included in Table 1 and here we show the results for the rest six patients in Table 3 .

It is worth noticing that for patient 26 and 40, with the scrubbing method we are unable to achieve the required level of forgetting even when the model generalisation on $\mathcal{D}_r$ and

Table 3: Same experiment as Table 1 with more patients. The patients are selected with different golden standard error on $\mathcal{D}_f$, from 0.12 for patient 59, up to 0.91 for patient 88.

Patient ID	Error on	Golden Standard	Scrubbing Method	Targeted Forgetting
88	$\mathcal{D}_r$	0.00	0.39	0.37
	$\mathcal{D}_f$	0.91	0.90	0.90
	$\mathcal{D}_{test}$	0.27	0.49	0.38
65	$\mathcal{D}_r$	0.01	0.36	0.33
	$\mathcal{D}_f$	0.74	0.75	0.75
	$\mathcal{D}_{test}$	0.26	0.47	0.44
77	$\mathcal{D}_r$	0.00	0.50	0.46
	$\mathcal{D}_f$	0.56	0.57	0.57
	$\mathcal{D}_{test}$	0.32	0.59	0.51
26	$\mathcal{D}_r$	0.00	0.79	0.39
	$\mathcal{D}_f$	0.42	0.00	0.42
	$\mathcal{D}_{test}$	0.19	0.75	0.57
40	$\mathcal{D}_r$	0.00	0.58	0.36
	$\mathcal{D}_f$	0.33	0.08	0.32
	$\mathcal{D}_{test}$	0.32	0.61	0.40
59	$\mathcal{D}_r$	0.00	0.67	0.22
	$\mathcal{D}_f$	0.12	0.09	0.11
	$\mathcal{D}_{test}$	0.22	0.61	0.38

$\mathcal{D}_{test}$ is seriously influenced. This means that the information scrubbed can hardly influence model's classification results, making these patients hard to forget with the scrubbing method. This further supports that these patients fall under the common cluster hypothesis. Overall, from Table 3, and in combination with our observations from Table 1, we conclude that the targeted forgetting can forget all patients, no matter one's data fall under edge case hypothesis so require more forgetting, or they belong to a cluster and are extremely hard to forget. Although sometimes our targeted forgetting brings relatively high error on $\mathcal{D}_{test}$, such as patient 77 (0.51) and patient 26 (0.57), there is still a small improvement compared with the scrubbing method: 0.59 for patient 77 and 0.75 for patient 26.

Appendix D. Implementation of patient-wise forgetting without golden standard model

Recall that our method adds noise to model weights highly informative to the forgetting data. During implementation, we are able to decide how many informative weights we would like to introduce noise to i.e. 1% most informative weights. When knowing a golden standard of forgetting, we usually affect less weights when forgetting an edge case patient to maintain good model generalisation error on $\mathcal{D}_{test}$, and affect more weights when forgetting a patient under the common cluster hypothesis to achieve the required level of forgetting.

For experiments in Section 3.3, without any prior knowledge of how much to forget and which hypotheses a patient falls from a golden standard, when implementing our method we cannot adapt the total number of weights being influenced for different cases. Therefore, we fix the number to 1% of total weights when applying the targeted forgetting. Note that 1% is a reasonable value selected based on extensive experiments. Here in Table 4 we additionally reported the value of the **High** level noise (defined in Table 2).

Table 4: The average noise value added to weights at High(1.00 error on $\mathcal{D}_f$)

Patient ID	Error on	Golden Standard	High Noise level	
			Scrubbing Method	Targeted Forgetting
94	$\mathcal{D}_r$ $\mathcal{D}_f$ $\mathcal{D}_{test}$	0.00 1.00 0.24	2.33E-05	3.00E-06
5	$\mathcal{D}_r$ $\mathcal{D}_f$ $\mathcal{D}_{test}$	0.00 0.81 0.25	1.65E-05	4.50E-06
13	$\mathcal{D}_r$ $\mathcal{D}_f$ $\mathcal{D}_{test}$	0.00 0.20 0.19	1.60E-05	8.66E-06
9	$\mathcal{D}_r$ $\mathcal{D}_f$ $\mathcal{D}_{test}$	0.00 0.01 0.23	1.43E-05	1.20E-05

From Table 4 we observe that to achieve the same forgetting on a patient, the scrubbing method requires adding more noise to model weights. This further explains why forgetting

with the scrubbing method in edge case particularly degrades the model generalisation on $\mathcal{D}_{test}$ compared with targeted forgetting: besides highly related weights, the scrubbing method also introduces large scale of noise to model weights that are with low relevance to the forgetting patient.