Why patient data cannot be easily forgotten?

Why patient data cannot be easily forgotten?

Ruolin Su[1*], Xiao Liu[1*], and Sotirios A. Tsaftaris[1,2]

[1] School of Engineering, University of Edinburgh, Edinburgh EH9 3FB, UK
[2] The Alan Turing Institute, London, UK
R.SU-1@ed.ac.uk, Xiao.Liu@ed.ac.uk, S.Tsaftaris@ed.ac.uk

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. However, 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. We show 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 dataset showcase the improved 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 [1,15]. Deep models can memorise a user's sensitive information [2,10,11]. Several attack types [23] including simple reverse engineering [7] can reveal private information of users. Particularly for healthcare, model inversion attacks can even recover a patient's medical images [24]. It is then without surprise why a patient may require that private information is not only deleted from databases but that any such information is forgotten by deep models trained on such databases.

A naive solution to forget a patient's data is to re-train the model without them. However, re-training is extremely time-consuming and sometimes impossible [21]. For example, in a federated learning scheme [17], 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.

As more advanced solutions, machine unlearning/forgetting approaches aim to remove private information of concerning data without re-training the model.

[*] Contributed equally

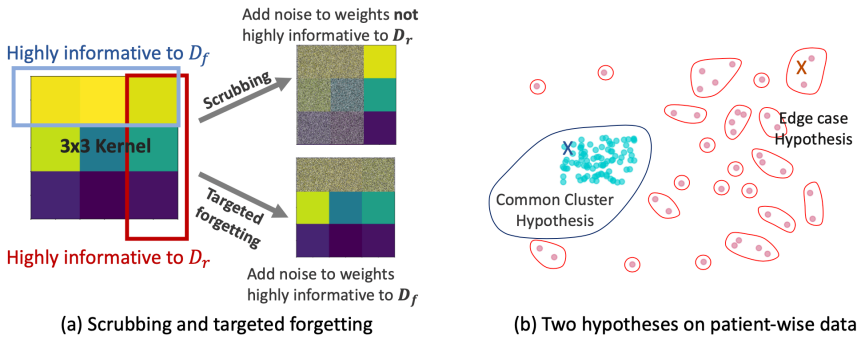

Fig. 1. (a) Visualisation of the scrubbing and targeted forgetting methods. $\mathcal{D}_r$ and $\mathcal{D}_f$ are the retaining data and the forgetting data. (b) Illustration of the two hypotheses. Blue contour delineates a big sub-population of similar samples within a *common cluster*; red contours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

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 data and made remarkable progress [5, 8, 9, 18, 19]. When the concept of machine unlearning/forgetting was first developed in [5], they discussed forgetting in statistical query learning [12]. Ginart et al. [8] specifically deal with data deletion in k-means clustering with excellent deleting efficiency. Another approach is to rely on variational inference and Bayesian models [18]. Recently, Sekhari et al. [19] propose a data deleting algorithm by expanding the forgetting limit whilst reserving the model's generalization ability. Golatkar et al. [9] address machine unlearning on deep networks to forget a subset of training data with their proposed scrubbing procedure (shown in Fig. 1(a)), which adds noise to model weights that are uninformative to the remaining data (training data excluding the concerning data) to achieve a weaker form of differential privacy [6].

Different from previous work, 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 patient. 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 [4, 16], 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.

We first study the patient-wise forgetting performance with simple translation of an existing machine unlearning method developed in [9]. For patients

[1] There is also a connection between edge cases and active learning [20], where one aims to actively label diverse data to bring more information to the model.

under different hypotheses, forgetting and generalisation performance obtained after scrubbing [9] vary as detailed in Section 3. In particular, the scrubbing method removes information not highly related to the remaining data to maintain good generalisation after forgetting, which is a weaker form of differential privacy [6]. When forgetting a patient under common cluster hypothesis, adequate performance can be achieved with the scrubbing method by carefully tuning the level of noise added to the model weights. 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. In fact, we observed in our experiment that data of a large portion of patients are edge cases while for computer vision datasets, the selected random cohort of data to be forgotten usually falls in the common cluster hypothesis. This limits the application of the scrubbing method and possibly other machine unlearning approaches that designed specifically for vision datasets to patient-wise forgetting.

To alleviate the limitation, we propose targeted forgetting, which only adds weighted noise to weights highly informative to a forgetting patient. In particular, we follow [9] 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 the proposed targeted forgetting, we can precisely forget edge case data and maintain good model generalisation performance. For patient data fall under the common cluster hypothesis, the algorithm can forget their information with the trade-off of the model performance on the whole cluster. This implies that for some patients within the common cluster hypothesis, it is not easy to forget them without negatively affecting the model.

Contributions:

1. We introduce the problem of patient-wise forgetting and formulate two hypotheses for patient-wise data.
2. We show that machine unlearning methods specifically designed for vision datasets such as [9] have poor performance in patient-wise forgetting.
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

Given a *training* dataset $\mathcal{D}$, a *forgetting* subset $\mathcal{D}_f \subset \mathcal{D}$ contains the images to be removed from a model $A(\mathcal{D})$, which is trained on $\mathcal{D}$ using any stochastic learning algorithm $A(\cdot)$. The *retaining* dataset is the complement $\mathcal{D}_r{=}\mathcal{D} \setminus \mathcal{D}_f$, thus $\mathcal{D}_r \cap \mathcal{D}_f = \emptyset$. Test data is denoted as $\mathcal{D}_{test}$. For patient-wise forgetting, $\mathcal{D}_f$ is all the images of one patient. Let $\mathbf{w}$ be the weights of a model. Let $S(\mathbf{w})$

denote the operations applied to model weights to forget $\mathcal{D}_f$ in the model, and $A(\mathcal{D}_r)$ be the *golden standard* model.

2.1 The scrubbing method

Assuming that $A(\mathcal{D})$ and $\mathcal{D}_r$ are accessible, Golatkar et al. [9] propose a robust scrubbing procedure modifying model $A(\mathcal{D})$, to brings it closer to a golden standard model $A(\mathcal{D}_r)$. They use FIM to approximate the hessian of the loss on $\mathcal{D}_r$, where higher values in FIM 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.

The Fisher Information Matrix F of a distribution $P_{x,y}(\mathbf{w})$ w.r.t. $\mathbf{w}$ defined in [9] is:

$$F = \mathbb{E}_{x\sim\mathcal{D}, y\sim p(y|x)}\left[\nabla_{\mathbf{w}} \log p_{\mathbf{w}}(y \mid x) \nabla_{\mathbf{w}} \log p_{\mathbf{w}}(y \mid x)^T\right] \tag{2}$$

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. In a medical dataset, the FIM ($F_{\mathcal{D}_r}(\mathbf{w})$) is derived by summing up the normalised FIM of each patient's data in the retaining set $\mathcal{D}_r$ and take the expectation at patient-level. Therefore, 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 weights corresponding to edge cases are relatively small.

A value within $F_{\mathcal{D}_r}(\mathbf{w})$ reflects to what extent the change to its corresponding weights $\mathbf{w}$ would influence the model's classification process on this set $\mathcal{D}_r$. 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. This also explains that weights correlated to data under common cluster hypothesis hold larger value than edge case hypothesis. Therefore, when scrubbing an edge case from a model, weights correlated to other edge cases even within $\mathcal{D}_r$ are also less informative to the remaining data thus will be scrubbed as well, making the model performance be negatively affected.

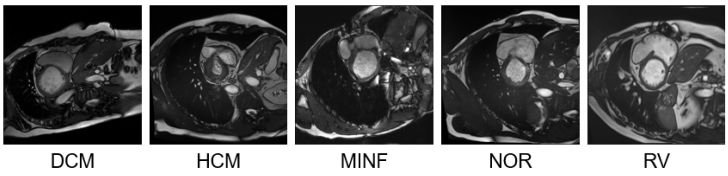

Fig. 2. Example images of ACDC dataset. DCM: dilated cardiomyopathy. HCM: hypertrophic cardiomyopathy. MINF: myocardial infarction. NOR: normal subjects. RV: abnormal right ventricle.

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 weights corresponding to $\mathcal{D}_r$ as in [9], our method precisely introduce noise to model weights highly informative about $\mathcal{D}_f$ (see Fig. 1(a)). Our proposed targeted forgetting is defined as:

$$\mathrm{S_T}(\mathbf{w}) = \mathbf{w} + \left(\lambda_T \sigma_{h_T}^2\right)^{\frac{1}{4}} F_{\mathcal{D}_f}(\mathbf{w})^{1/4}, \tag{3}$$

where λ_T and σ_{h_T} are analogous parameters to λ and σ_h defined in Eq 1.
Performance on the two hypotheses *Common cluster hypothesis*: Targeted forgetting will add noise to the most informative model weights corresponding to $\mathcal{D}_f$ so it will also 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 an edge case and maintain good model performance. Results and discussion are detailed in Section 3.

3 Experiments

We first explore why scrubbing [9] works well on computer vision datasets but shows poorer performance on patient-wise forgetting. We conduct an experiment to demonstrate the intrinsic dataset biases of CIFAR-10 [14] and ACDC [3]. 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 224×224 2D images, there are 14,724 images from 90 patients

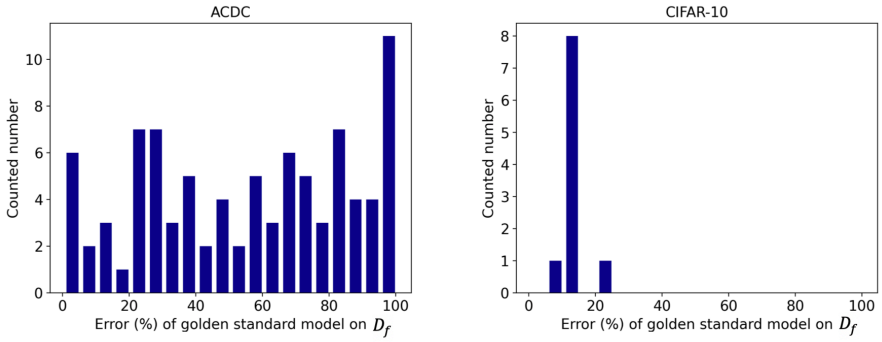

Fig. 3. Histograms of re-training experiments. The y-axis refers to the total number of patients/sets whose golden standard lies within an interval(e.g. [95,100]) of x-axis.

form D, and 1,464 images from 10 patients form $\mathcal{D}_{test}$. 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 [9]. When training the ACDC classifier, the model has a VGG-like architecture as in [22]. We use Cross Entropy as the loss function and use Adam optimizer [13] 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 $\mathcal{D}_{test}$.

3.1 The hardness of patient-wise forgetting

Here we compare between CIFAR-10 and ACDC to show that some patient data are hard to learn and forget. For 90 patients in ACDC, we remove one patient's data and re-train a model on the remaining 89 to be the golden standard model, $A(\mathcal{D}_r)$. We then measure the error of the deleted patient on $A(\mathcal{D}_r)$. 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 re-train experiments. Data are hard to generalize by its golden model show high error on $A(\mathcal{D}_r)$ and thus should be hard 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 classification error of a $\mathcal{D}_f$ on its corresponding golden model $A(\mathcal{D}_r)$ 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;

Table 1. Forgetting results for four patients. We report Error = 1−Accuracy on the forgetting ($\mathcal{D}_f$) and test ($\mathcal{D}_{test}$) sets respectively. Scrubbing Method refers to the method of [9] whereas Targeted Forgetting refers to the method in Section 2.2. Red and blue denote the golden standard of forgetting performance for each row respectively, with performance being better when it is closer to the standard. 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. The confidence bar is obtained over three experiments.

Patient ID	Error on	Golden Standard	Noise level					
			Low		Medium		High	
			Scrubbing	Targeted Forgetting	Scrubbing	Targeted Forgetting	Scrubbing	Targeted Forgetting
94 (Edge)	$\mathcal{D}_f$	1.000±0.000	0.154±0.005	0.174±0.020	0.859±0.010	0.851±0.018	1.000±0.000	1.000±0.000
	$\mathcal{D}_{test}$	0.237±0.002	0.671±0.012	0.223±0.011	0.739±0.007	0.291±0.005	0.746±0.008	0.316±0.002
5 (Edge)	$\mathcal{D}_f$	0.809±0.009	0.127±0.022	0.121±0.019	0.853±0.020	0.857±0.002	0.997±0.003	1.000±0.000
	$\mathcal{D}_{test}$	0.253±0.026	0.394±0.017	0.269±0.004	0.624±0.015	0.407±0.001	0.696±0.002	0.506±0.002
13 (Cluster)	$\mathcal{D}_f$	0.202±0.004	0.111±0.006	0.092±0.002	0.871±0.018	0.850±0.021	1.000±0.000	1.000±0.000
	$\mathcal{D}_{test}$	0.194±0.012	0.361±0.001	0.343±0.007	0.590±0.005	0.524±0.013	0.694±0.004	0.602±0.016
9 (Cluster)	$\mathcal{D}_f$	0.010±0.002	0.176±0.005	0.152±0.009	0.892±0.003	0.862±0.005	0.998±0.002	0.995±0.005
	$\mathcal{D}_{test}$	0.233±0.007	0.402±0.012	0.442±0.001	0.643±0.006	0.613±0.001	0.699±0.005	0.656±0.001

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**. This is remarkably different in CIFAR-10: golden model results concentrate at low error indicating that few edge cases exist. In addition, as discussed in section 2.1, when dealing with edge cases, scrubbing can degrade model performance. This will explain the results of the scrubbing method: under-performance in ACDC because many patients fall under the edge case hypothesis.

3.2 Patient-wise forgetting performance

We focus on four representative patients using the analysis in Section 3.1: patients 94 and 5 that fall under the edge case hypothesis; and patients 13 and 9 fall under a common cluster hypothesis. Here we consider a stringent scenario: the re-trained golden standard model is not available for deciding how much to forget, so the level of noise to be added during scrubbing or forgetting is unknown. We adjust noise strength (low, medium and high) by modulating the hyperparameters in both methods to achieve different levels of forgetting.[2] 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 $\mathcal{D}_{test}$.

Is targeted forgetting better for forgetting edge cases? For edge cases, forgetting can be achieved (compared to the golden standard) at high level of

[2] For our experiments we fix to introduce noise to 1% most informative weights (based on extensive experiments) when applying the targeted forgetting.

Table 2. The average noise value added to weights at High (when achieving 1.00 error on $\mathcal{D}_f$). Note that medium and low noise is with 66.7% and 30.0% of high noise level respectively.

Patient ID	High	
	Scrubbing	Targeted Forgetting
94 (Edge)	2.33E-05	3.00E-06
5 (Edge)	1.65E-05	4.5E-06
13 (Cluster)	1.6E-05	8.66E-06
9 (Cluster)	1.43E-05	1.2E-05

noise with both methods. However, the scrubbing method significantly degrades the model generalisation performance. With targeted forgetting, good model generalisation performance on $\mathcal{D}_{test}$ at all noise levels is rather maintained. Additionally from Table 2 we observe that the scrubbing method adds more noise to model weights to forget an edge case. This further supports our discussion in section 2.1 on how the scrubbing method negatively affects the overall model performance when forgetting edge case.

Is targeted forgetting better for forgetting common cluster cases? For common cluster cases, both methods can achieve standard forgetting with a near low level of noise with nice model's generalisation performance on $\mathcal{D}_{test}$, as shown in Table 1. For example for patient 13, the test error of two methods at low noise level is 0.361 and 0.343, which is close and relatively small. When the noise level grows to medium and high to forget more, although the test error with two methods still being close, it grows to a high value. Overall, when forgetting common cluster cases, the two methods show similar good performance at a standard level of forgetting and they both can forget more about a patient by sacrificing the model's generalisation.

Can patient data be completely forgotten? Overall, for edge cases, using targeted forgetting, the patient-wise data can be completely forgotten (achieving error higher than 0.80 (random decision for 5 classes in our case) on $\mathcal{D}_f$) without sacrificing the model generalisation performance. While for common cluster cases, it is less likely to forget the patient data as completely forgetting will result the significantly degraded generalisation performance with the scrubbing or our targeted forgetting. In fact, the level of noise added to the model weights affects the trade-off between model performance and respecting data protection. Higher noise leads to more information being removed, thus protecting the data better yet degrading the model accuracy. Therefore, the noise needs to be carefully designed such that a sweet spot between forgetting and generalisation performance can be achieved.

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.

5 Acknowledgements

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 RC-SRF1819\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

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Arpit, D., Jastrzkebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., 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)
4. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018)
5. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: 2015 IEEE Symposium on Security and Privacy. pp. 463–480 (2015). `https://doi.org/10.1109/SP.2015.35`
6. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3-4), 211–407 (2014)
7. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1322–1333 (2015)

8. Ginart, A., Guan, M.Y., Valiant, G., Zou, J.: Making ai forget you: Data deletion in machine learning. arXiv preprint arXiv:1907.05012 (2019)
9. Golatkar, A., Achille, A., Soatto, S.: 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)
10. Hartley, J., Tsaftaris, S.A.: Unintended memorisation of unique features in neural networks. arXiv preprint arXiv:2205.10079 (2022)
11. Jegorova, M., Kaul, C., Mayor, C., O'Neil, A.Q., Weir, A., Murray-Smith, R., Tsaftaris, S.A.: Survey: Leakage and privacy at inference time. arXiv preprint arXiv:2107.01614 (2021)
12. Kearns, M.: Efficient noise-tolerant learning from statistical queries. Journal of the ACM (JACM) **45**(6), 983–1006 (1998)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In proc. ICLR (2015)
14. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009)
15. Liu, X., Tsaftaris, S.A.: 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. pp. 95–105. Springer (2020)
16. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2537–2546 (2019)
17. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
18. Nguyen, Q.P., Low, B.K.H., Jaillet, P.: Variational bayesian unlearning. Advances in Neural Information Processing Systems **33** (2020)
19. Sekhari, A., Acharya, J., Kamath, G., Suresh, A.T.: Remember what you want to forget: Algorithms for machine unlearning. arXiv preprint arXiv:2103.03279 (2021)
20. Settles, B.: Active learning literature survey (2009)
21. Shintre, S., Roundy, K.A., Dhaliwal, J.: Making machine learning forget. In: Annual Privacy Forum. pp. 72–83. Springer (2019)
22. Thermos, S., Liu, X., O'Neil, A., Tsaftaris, S.A.: Controllable cardiac synthesis via disentangled anatomy arithmetic. In: MICCAI (2021)
23. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing (2019)
24. Wu, M., Zhang, X., Ding, J., Nguyen, H., Yu, R., Pan, M., Wong, S.T.: Evaluation of inference attack models for deep learning on medical data. arXiv preprint arXiv:2011.00177 (2020)

6 Appendix

We explore if overfitting would be an issue affecting the results in Section 3.1 by redoing the experiment by early-stop training models. With all the settings being the same as in Section 3, the training epochs for the 90 individual models is changed from 13 to 7 to obtain less overfitted models.

Fig. A1 collects the results with early stop models. Overall, compared with Fig.3 in Section 3.1, although the distribution of the early stop results histogram

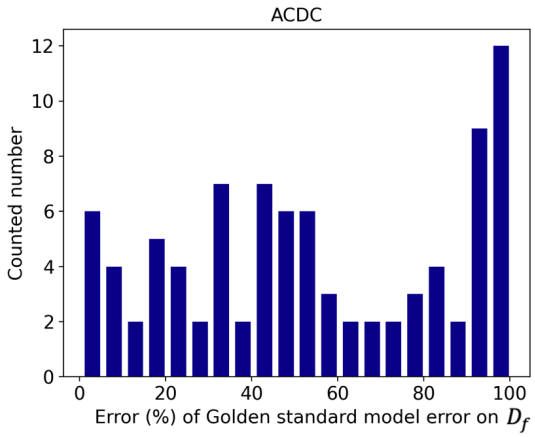

Fig. A1. Histograms of re-training experiments. The y-axis refers to the total number of patients/sets whose golden standard lies within an interval(e.g. [95,100]) of x-axis.

is slightly different, the 90 individually measured results of classification error of a $\mathcal{D}_f$ on its corresponding golden model $A(\mathcal{D}_r)$ also vary from 0% to 100%. By considering a threshold of 50% on the error of the golden model, there are still $> 50\%$ of patients in ACDC can be considered to belong to the edge case hypothesis. Therefore, overfitting is not considered the reason for the emergence of edge cases.